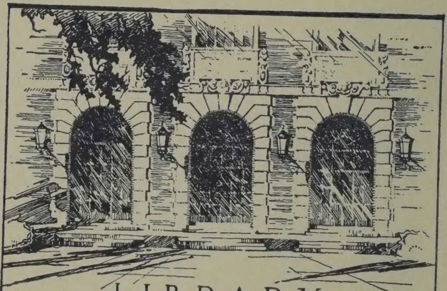


LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

[redacted]

Allerton Conference on Circuit
and System Theory 1966

Proceedings

720
V.4

Proceedings

FOURTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

October 5-7, 1966
University of Illinois
Allerton House, Monticello, Illinois

Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois, Urbana, Illinois and the
Circuit Theory Group of the
Institute of Electrical & Electronics Engineers

DIGITAL COMPUTER
LABORATORY
LIBRARY

OCT 25
1968

DIGITAL COMPUTER
LABORATORY
LIBRARY

PROCEEDINGS
FOURTH ANNUAL ALLERTON CONFERENCE
ON CIRCUIT AND SYSTEM THEORY

EDITED BY

WILLIAM R. PERKINS
J. B. CRUZ, JR.

Conference Co-Chairmen

Conference held

October 5, 6, 7, 1966
Allerton House
Monticello, Illinois

Sponsored by the
DEPARTMENT OF ELECTRICAL ENGINEERING
and the
COORDINATED SCIENCE LABORATORY
of the
UNIVERSITY OF ILLINOIS
Urbana, Illinois

and the

CIRCUIT THEORY GROUP
of the
INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS

FOREWORD

The Fourth Annual Allerton Conference on Circuit and System Theory was held at Allerton House from October 5 to 7, 1966. These Proceedings are a permanent record of the papers presented at that Conference. Again, as in the case of the three previous Allerton Conferences, the theme was new advances in circuit and system theory, with special emphasis on the interrelation of these two fields.

We acknowledge with pleasure the partial support of the sensitivity sessions given by the Air Force Office of Scientific Research under Grant AFOSR 931.66, and the cooperation of Colonel B. R. Agins. The Conference was co-sponsored by the University of Illinois and the Circuit Theory Group of the Institute of Electrical and Electronics Engineers.

We acknowledge the encouragement and support of Professor W. Dale Compton, Director of the Coordinated Science Laboratory, and Professor E. C. Jordan, Head of the Department of Electrical Engineering at the University of Illinois. Special thanks are due Mrs. Divona Keel, who coordinated with extraordinary skill all Conference arrangements. Her assistance was indispensable in making the Conference successful. We thank Mr. Harold B. Lawler and Mrs. Rosa Townsend for their help with certain business aspects of the Conference. We are indebted to our colleagues for their vital assistance: Professors Robert T. Chien, Abraham Haddad, Gernot Metze, Wataru Mayeda, Franco Preparata, Timothy Trick, and Nelson Wax.

W. R. Perkins
J. B. Cruz, Jr.

TABLE OF CONTENTS

Foreword iii

GENERAL SESSION A 1
Chairman: W. R. Perkins, University of Illinois

On Global Sensitivity,
Mihajlo D. Mesarovic and Yasuhiko Takahara 1

SENSITIVITY I 8
Chairman: G. J. Thaler, U. S. Naval Postgraduate School

On the Compatibility of Adaptive Controllers,
S. P. Bingulac 8

The Application of Sensitivity Methods to the Control of Adaptive Systems,
Agostino Burzio and Richard C. Dorf 17

A Transformation Improving Stability of Adaptive Systems,
S. P. Bingulac 28

TIME-VARYING SYSTEMS 36
Chairman: T. N. Trick, University of Illinois

A Transform Technique for Time-Varying Linear Hybrid Systems,
Frederick M. Waltz 36

Approximate Analysis of Linear Time-Varying Systems - A Quantitative Approach
N. DeClaris and H. Van Landingham 46

Estimation of the Impulse Response of Time-Varying Linear Systems,
William M. Hammond and George R. Cooper 53

NETWORK ANALYSIS 61
Chairman: W. Mayeda, University of Illinois

An Algebraic Manipulation Method for Network Analysis,
Yasutoshi Ishizaki, Tohoru Tsuchiya, Nobuyoshi Yoshida 61

Natural Properties of Cascaded Identical Symmetric Networks,
Joseph Wood Rogers 70

Linear Degenerate RC-Network,
Sik-Kee Au and Ruey-Wen Liu 71

SYNTHESIS OF 1-PORT NETWORKS 72
Chairman: D. Calahan, University of Michigan, Ann Arbor

Design of One-Ports Using Only Resonators,
John F. O'Neill 72

Synthesis of Driving Point Impedances with Reactive Gyrators,
Elmer A. Hoyer and Gladwyn V. Lago 82

Classifications and Patterns of Recognition for Positive Real Immittance Functions,
Kurt H. Haase 83

PATTERN RECOGNITION AND LEARNING 90
Chairman: R. T. Chien, University of Illinois

A Limit Theorem with Pattern Recognition Applications,
Terry J. Wagner 90

Performance and Implementation of the k-Nearest Neighbor Decision Rule with Incorrectly Identified Training Samples,
A. W. Whitney and S. J. Dwyer III 96

On Bayesian Learning and Stochastic Approximation,
Y. T. Chien and K. S. Fu 107

FILTERS I 116
Chairman: D. Calahan, University of Michigan, Ann Arbor

The Synthesis of Cascade Lattice Crystal Filters,
James D. Schoeffler and Allan D. Waren 116

An Asymmetrical Lattice for Crystal Band-Pass Filters,
John A. C. Bingham 141

Broad-Band Matching by Lowpass Transformations,
E. Christian and E. Eisenmann 155

Explicit Formulas for Butterworth and Chebycheff Ladder Filters with New Patterns of Reflection Zeros,
I. Navot, B. Gonen and Z. Peri 165

Design of Numerical Filters: Scanning Functions and Equal Ripple Approximation,
G. E. Heyliger 175

Equiripple Approximations of Group Delay and Attenuation in Network Synthesis,
R. K. Kwan and G. G. Bach 186

TOPOLOGICAL NETWORK ANALYSIS AND SYNTHESIS 197
Chairman: W. Mayeda, University of Illinois

A Topological Approach to the Analysis of Linear Circuits,
S. L. Hakimi and N. Deo 197

On the Generation of Non-Singular Submatrices and Their Corresponding Subgraphs,
Wai-Kai Chen 207

A Graph-Theory Approach to the Analysis of Multiterminal Element Networks,
Mirko M. Milic 218

A New Topological Technique for Solving a Linear System,
S. Park Chan and Hsia-Nung Mai 227

Topological Formulas for General Linear Networks,
Maw-Tsuey Jong and George W. Zobrist 237

Network Topology and Stability of Nonlinear RLC Networks,
Tadao Murata 244

Digital Computer Synthesis of Admittance Matrices of N+1 Nodes,
Elmer A. Hoyer, George Perdikaris, and George W. Zobrist 249

SENSITIVITY II 250
Chairman: M. E. Van Valkenburg, Princeton University

Frequency Sensitivity of Active Networks,
E. C. Ball and G. J. Thaler 250

Worst Case Gain Sensitivity With Zero Phase Sensitivity in Active RC Network Synthesis,
H. E. Jones and B. A. Shenoi 259

Sensitivity Minimization in Active RC Integrated Circuit Design,
Samuel C. Lee 269

A New Definition of Multiparameter Sensitivity,
R. N. Biswas 282

Eigenfunction Sensitivity and Invariance,
D. C. Reddy 283

STOCHASTIC SIGNALS AND SYSTEMS 291
Chairman: R. B. Ash, University of Illinois

Characterization of a Simple Nonlinear Systems,
M. L. Liou 291

On the Inclusion Probability of a Stochastic Dynamical System,
Leo J. Van Mellaert 300

Passage of Stochastic Input Through a Class of Nonlinear Systems,
M. Arozullah and G. S. Glinski 310

Integral Representations for Nonuniform Sampling Expansions,
A. H. Haddad and J. B. Thomas 322

On the Selection of Optimum Exponents for the Series Representation of Certain Stochastic Processes,
Arthur J. Carlson and William D. Wade. 334

SEQUENTIAL SWITCHING AND INFORMATION SYSTEMS 335
Chairman: F. Preparata, University of Illinois

Detecting Races, Hazards and Oscillations in Logical Systems,
E. L. Calvin 335

Generation of an Output Regular Expression of a Sequential Machine with a Specified Input Regular Expression,
S. S. Yau 343

A Queuing Analysis of the Number of Messages in a Store and Forward Communication Network,
Stephen D. Shapiro 352

Topological Structures of Information Retrieval Systems,
R. T. Chien and F. P. Preparata 359

MEETING

Joint meeting with the Chicago Area Chapter of the IEEE Group on Information Theory

Use of Arithmetic Codes in Computers,
D. E. Muller 368

GENERAL SESSION B 377
Chairman: J. B. Cruz, Jr., University of Illinois

Development and Use of Sensitivity Models,
P. Kokotovic, J. Medanic and R. Rutman 377

Spectral Factorization Using the Riccati Equation,
B. L. Ho and R. E. Kalman 388

Development and Applications of a System Theory Criterion for Rational Positive Real Matrices,
B. D. O. Anderson 400

GRAPH THEORY 408
Chairman: M. E. Van Valkenburg, Princeton University

Number of Trees in a Cascade of Two-Port Networks,
B. R. Myers 408

Minimal Realization of a Class of Max-Flow Min-Outage Weighted Nets,
B. R. Myers and M. A. Tapia 416

Analysis and Synthesis of Node Weighted Networks from the Simultaneous Flow Standpoint,
Enrique A. Davila 426

A Class of Planar Graphs Containing Hamiltonian Circuits,
D. T. Tang 427

Additional Simple Codes for Planar Graphs,
Louis Weinberg 433

Alternative Algorithms for Determining Isomorphisms and Automorphism Groups of Planar Graphs,
Louis Weinberg 444

Algorithms for theDichotomous Representation of Macrocircuits,
R. M. Carpenter and W. W. Happ 454

FILTERS II 465
Chairman: B. A. Shenoi, University of Minnesota

Maximally Flat Approximation Techniques,
H. J. Orchard and G. C. Temes 465

A General Class of Tunable Active R-C Filters,
J. V. Leeds, Jr., ... 466

A Novel Approach in the Design of Audio Phase Interference Filters Using Highly Damped Circuit,
Lawrence V. Kriger 474

On a Class of Finite-Memory Digital Filters,
George Shuraym 485

Insertion-Loss Analysis and Exact Design of the Equal-Element Class of TEM Transmission-Line Filters,
Harvey J. Hindin and Jesse J. Taub 486

The Synthesis of Orthogonal Filters by the Use of the Sturm-Liouville Equations,
Micheal Parten and R. H. Seacat 487

Morgan-Voyce Polynomials in Recurrent Ladders,
M. N. S. Swamy and B. B. Bhattacharyya 488

SENSITIVITY III 489
Chairman: Major Bernard S. Morgan, The Frank J. Seiler Research Laboratory

Trajectory Sensitivity of an Optimal Control Systems,
Henry D'Angelo, Maynard L. Moe and Thomas C. Hendricks 489

Low Sensitivity Design of Optimal Linear Control Systems,
Stephen J. Kahne 499

Sensitivity of Optimal Control Systems,
N. K. Sinha and Satya Ratnam Atluri 508

A Design Technique for Improving System Sensitivity,
William A. Porter 517

Synthesis for Minimum Sensitivity Under Worst Case Conditions,
Richard Lawrence Gonzales 527

Sensitivity Integrals,
Sydney R. Parker 538

Design of Multivariable Linear Systems by Minimization of Parameter Induced Error,
James G. Simes and George A. Etzweiler 548

OPTIMAL CONTROL 549
Chairman: B. C. Kuo, University of Illinois

An Application of Bounded Phase-Coordinate Control Theory to an Undamped Oscillatory Plant with Control Amplitude and Rate Saturations,
J. Y. S. Luh and J. S. Shafran 549

Optimization and Implementation of System Control Laws,
M. Schoenberger 557

On the Quadratic Control Problem with an Energy Constraint,
Gerald Cook and James E. Funk 567

Asymptotic Solution of the Matrix Riccati Equation of Optimal Control,
John J. O'Donnell 577

Discrete Linear System with Convex Cost Functional,
Dong Hak Chyung 587

A Sufficiency Condition for Minimum Variance Control of Markov Processes,
M. K. Sain 593

Optimum Combined Control and Estimation for Partially Controlled Systems,
Robert E. Larson 600

Locally Optimum Differential Games,
Celso de Renna e Souza 611

NETWORK THEORY 621
Chairman: M. Sobral, Jr., University of Michigan, Dearborn

Signal Processing by Passive Linear Networks,
James A. Resh 621

State Variables, Energy Functions, and PR Matrices,
B. J. Leon and P. M. Lin 630

State Representations, Passivity, Reciprocity, and N-Port Synthesis,
David M. Layton 639

A Canonical Network for Periodically Variable Linear Systems,
H. E. Meadows, L. M. Silverman, and L. E. Franks . . 649

Stability of Linear Time-Varying Networks,
K. R. Kleckner and R. C. Waag 659

Constant Resistance, Wide-Sense Solvability, and Self-Duality,
C. A. Desoer, R. W. Newcomb, and K. K. Wong 669

Compatibility of Immittance Matrices,
S. K. Mullick 674

A Numerical Method for Determining Nodal Voltages in D. C. Circuits with Nonlinear Elements,
R. L. Duty, J. B. Eggerling, and C. G. Hooks 684

GENERAL SESSION C 685
Chairman: W. R. Perkins, University of Illinois

Path Integrals, Liapunov Functions, and Quadratic Minimization,
R. W. Brockett 685

ACTIVE NETWORKS. 698
Chairman: B. J. Leon, Purdue University

The Rotator - A New Linear Active 2-Port Network Element,
Leon O. Chua 698

A New Method for the Determination of Gain in Parametric Amplifiers,
J. V. Adams and D. R. Anderson 708

Subharmonic Oscillations in Parametric Circuits,
T. N. Trick 718

On the Limitations Governing Operation of Bipolar and Field-Effect Transistors in Circuits and Systems,
Keats A. Pullen, Jr. 726

Design Theory of a Negative Resistance Amplifier Using a Transistorized Isolator,
R. S. Gopalan and B. A. Shenoi 731

COMPUTER-ORIENTED SYSTEMS TECHNIQUES 741
Chairman: S. Kahne, University of Minnesota

Identification of Nonlinear Sampled-Data Systems,
K. Steiglitz and L. E. McBride 741

Self-Oscillation of Nonlinear Systems with Gaussian Noise,
S. H. Wu and J. A. Heinen 751

Analysis of Delta-Modulated Feedback Systems,
S. H. Wu and A. N. Michel 752

The Conjugate Gradient Minimization Method in Function Space,
A. D. Waren, L. S. Lasdon, and S. K. Mitter 753

Time-Domain Synthesis by Nonlinear Programming,
I. A. Athanassopoulos, J. D. Schoeffler and A. D. Waren 766

SYSTEM THEORY 776
Chairman: R. W. Brockett, Massachusetts Institute of Technology

Equivalence and Synthesis of Time-Variable Linear Systems,
L. M. Silverman and H. E. Meadows 776

Controllability and Observability of Composite Systems,
C. T. Chen and C. A. Desoer 785

On Relationships Between the Theories of Discrete-State and Continuous-State Linear Time-Invariant Systems,
P. H. Swain and K. S. Fu 794

Hilbert Transforms and Generalized Functions,
Harry Schwarzlander 805

STABILITY I 814
Chairman: T. N. Trick, University of Illinois

A Conjecture on the Existence of a Quadratic Liapunov Function for Linear Time-Invariant Feedback Systems,
Kumpati S. Narendra and Charles P. Neuman 814

The Network Analog as a Heuristic Tool for Generating Stability Criteria,
Thomas E. Stern 819

On a Generalization of the Circle Criterion,
Michael Gruber and Jacques L. Willems 827

Perturbation Theory for the Analysis of Instability in Nonlinear Feedback Systems,
Jan C. Willems 836

DISTRIBUTED DEVICES AND SYSTEMS 849
Chairman: B. A. Shenoi, University of Minnesota

Reciprocal Time Domain Analysis of Distributed Networks,
W. W. Happ and Someshwar C. Gupta 849

A Frequency Transformation for Commensurate Transmission Line Networks,
Edward G. Cristal 858

Variable Parameter Transmission Line Theory,
E. C. Bertnolli, Yu Kuo Chen, and C. A. Halijak . . 868

NONLINEAR SYSTEMS 869
Chairman: G. H. Fett, University of Illinois

Estimating the Transient Response of Nonlinear Systems,
Gary L. Blank and Larry J. Levy 869

Performance Investigation of Nonlinear Systems by Observing Eigenvalue Movement,
R. Wasta and S. A. Steele 876

System Performance of Control Systems with Two Nonlinearities,
C. A. Pellegrini, H. M. Yockey, and G. J. Thaler . 877

Parameter Plane Analysis of Forced Nonlinear Oscillations,
J. Moore and D. Siljak. 887

LINEAR CONTROL SYSTEMS 903
Chairman: B. C. Kuo, University of Illinois

Identification of a Linear System from Discrete Values of Input-Output Data,
C. N. Weygandt and N. N. Puri 903

Response of Sampled-Data-Hold Systems Between Sampling Instants,
Willie L. McDaniel, Jr. 913

An Approximate Identity Operator for Continuous Servo-Mechanisms with Time Lag,
Glen H. Fountain and Charles A. Halijak 914

STABILITY II 915
Chairman: N. Wax, University of Illinois

On the Absolute Stability of Systems Containing Multiple Nonlinearities,
M. D. Srinath, M. A. L. Thathachar and H. K. Ramapriyan 915

A Generalized Zubov Formulation for Limit Cycles in Third Order Nonlinear Systems,
M. J. Wozny and D. G. Schultz 916

Input and Output Bounds for Forced Nonlinear Control Systems,
Rui J. P. de Figueiredo and Michel M. De Vries . . . 925

A Study of System Behavior for Nonnormal Systems Near and at Infinity of the Phase Plane,
Wayne H. Chen 933

NETWORK SYNTHESIS 934
Chairman: J. B. Cruz, Jr., University of Illinois

Network Synthesis Using Lossy Reactances,
Jonny Andersen and Harry B. Lee 934

Primitive Matrices for Doubly Terminated Ladder Networks,
Thomas G. Marshall, Jr. 935

Realization of the Equilibrium and Response Immittances of Transformerless RLC Networks,
J. E. Olivares, Jr. 944

Non-Series-Parallel Realization of Bisymmetrical Two-Element-Kind Two-Ports,
Samuel C. Lee 955

ON GLOBAL SENSITIVITY

MIHAJLO D. MESAROVIC AND YASUHIKO TAKAHARA
Systems Research Center - Case Institute of Technology
Cleveland, Ohio

INTRODUCTION

Concept of sensitivity is introduced in reference to the behavior of a system in a class of variable conditions. Customarily, the sensitivity is defined on a local basis, i.e., starting from some given conditions the sensitivity of the system reflects the changes in the behavior of the system when the conditions change. The concept of sensitivity introduced here will refer to the behavior of the system in an entire class of variable conditions and will be termed the global sensitivity. To arrive at the notion of the global sensitivity it is necessary to introduce a way for comparing (or coordinating) the behavior of the system. This will be done here via so-called satisfaction problem, (1),(2),(3) which will be presented as being intimately related with the global sensitivity.

AN EXAMPLE AND A CLASS OF PROBLEMS

To indicate the types of problems which are responsible for the introduction of the concept of global sensitivity, some examples will be considered first.

Consider a system $S: M \times U \to Y$, described by the equation, $y = 2m + u$, where $y \in Y$, $m \in M$, $u \in U$ and Y, M, U are subsets of real numbers. M is the set of controls (decisions, manipulated variable) while U is the set of disturbances. Given a positive real c and let $U = \{c, -c\}$, $M = \{m: -c \leq m \leq c\}$. Let the performance of the system be evaluated by the absolute value of the output y, i.e. $q = \| y \|$.

An element of M can be selected at will and when it is applied the behavior of the system is determined solely by the conditions specified by the second input $u \in U$. Primary concern in a reasonable engineering design (or synthesis) is to have a "tolerable" performance under any of these conditions, i.e. for any $u \in U$. Selection of a single $m \in M$ can ensure

---

*This paper was supported by ONR Contract 1141(12).

a prespecified performance only for a single u, say $u_m \in U$, and it is necessary then that the performance does not significantly change over the entire U, i.e. remains close to the performance for u_m. That, of course, implies that the system performance should be significantly insensitive over the set U. The problem then arises from the fact that the very procedure for the selection of $m \in M$ can increase the sensitivity to such an extent that the performance becomes unsatisfactory for some $u \in u$.

For the example under consideration, assume that the specifications require that the performance does not exceed c, i.e. $q \leq c$ for any $u \in U$. In a standard procedure one would select a $u_1 \in U$ and find $m_1 \in M$ such that q is minimized. If $u_1 = c$ the optimal m is $m_1 = -\frac{1}{2}c$ leading to $q = 0$. The performance for $u_2 = -c$ is then $q = 2c$ i.e. outside the limits. For another possibility, $u_2 = -c$, the optimal m is $m_2 = \frac{1}{2}c$ and the performance for $u_1 = c$ becomes $q = \| -2c \| = 2c$; again outside the limits. The question then arises whether there exists $m \in M$ which is not optimal for any $u \in U$ yet it yields the performance of the system within the tolerance limits (i.e. below c) for all $u \in U$. Such m_s does exist; $m_s = 0$ leads to $q = c$ for both u_1 and u_2.

The situation from the example is quite general. For instance, let X be a normed linear space and the system be described by the equation $y = F(m) + u$ where F is an operator (not necessarily linear) such that $F: X \rightarrow X$: let $M = \{x: \ x \ \leq c\}$ and U is the boundary set for M. Finally, let the performance be $q(m,u) = \| y \| = \| F(m)+u \|$ and the specifications require that the performance does not exceed c under any conditions. If the operator F is such that: (i) $F(m) = o$ if and only if $m = o$; (ii) for any $u \in U$ there exists $m_u \in M$ such that $\inf \| F(m_u) + u \| < c$; the performance of the system exceeds the limit (i.e. c) for any $m \in M$ which is optimal with respect to a given $u \in U$. Yet the performance remains within limits for $m = 0$ which is not optimal with respect to any $u \in U$(4).

The conclusion from these examples is that if there exists a variety of conditions for the behavior of the system the problem of selecting m should be defined so that the prescribed performance is achieved over the entire U. This situation leads to the concept of global sensitivity.

GLOBAL SENSITIVITY

Let the system S be a mapping, $S: M \times U \rightarrow Y$, and the behavior of S is evaluated by the performance function $G: M \times Y \rightarrow Q$ (For simplicity of notation we shall assume that U represents both the disturbance and the difference in the initial states) For any given $m \in M$ the system becomes $S = S|\{m\} = \{m\} \times U \rightarrow Y$, i.e., it is completely specified by $S_m: U \rightarrow Y$. Similarly, for any $m \in M$ the performance function becomes $G_m: U \rightarrow Q$. Global sensitivity then is a set function T mapping power set of Q, (denoted by $\pi(Q)$), into a linearly ordered set Z, i.e. $T: \pi(Q) \rightarrow Z$. With any $m \in M$ there is associated a subset $Q_m \subset Q$ such that $Q_m = G_m(U)$ and also an element of Z, i.e., $z_m = T(Q_m)$. Linear ordering relation in Z (denoted by $<_z$) induces a linear ordering in M (denoted $<_m$,) such that $m <_m m' \leftrightarrow T(Q_{s_m}) <_z T(Q_{s_{m'}})$. A system $S_{m'}$ will be called less sensitivite than another system S_m if $m <_m m'$.

For further discussion in this paper it will be assumed that: (i) Z is the set of real numbers, (ii) U is a measurable space; (iii) there is given a measurable function $W_m: U \rightarrow Z$; (iv) the set function T is defined by the integral $z_m = \int_U W_m(u)\, d\ (u)$ where μ is the measure in U.

Global sensitivity function T depends then upon the selection of W_m. Some useful selections of W_m are based on the so-called satisfaction problem.(1)(3) Assume that in addition to S and G there is given a function $V: U \rightarrow Q$ and Q is linearly ordered by $<_q$. For any given $u \in U$ the system behavior will be called satisfactory if $G_m(u) \leq_q V(u)$. V is termed tolerance function since for any $u \in U$, $V(u)$ serves as a tolerance limit for the performance. Two typical selections for W_m are then: (i) W_m is the characteristic function for the subsets in U, i.e. given any $U' \subset U$, $W_m(u) = 1$ for any $u \in U'$ and $W_n(u) = 0$ for any $u \notin U'$. To any $m \in M$ corresponds a subset in the disturbance space $U_m \subset U$ such that the performance is satisfactory in U_m, i.e. $u \in U_m$ if and only if $G_m(u) \leq V(u)$. Since U_m is measurable $z_m = u(U_m)$, i.e. the global sensitivity is the measure of the subset U_m. A system S_m will be then less sensitive than another system $S_{m'}$, if its performance is within the tolerance limit in a larger disturbance set (in the sense of the measure μ). (ii) Another useful selection is $W_m(u) = V(u) - G_m(u)$, i.e. $W_m(u) = V(u) - G[F(m,u),u]$.

OPTIMAL DOMAIN AND SATISFACTION DOMAIN

In order to consider the role of global sensitivity in the synthesis, (in particular, based on the optimization procedures) we shall define the following three sets: (i) Assume that an optimization procedure is given so that for any $u \in U$ there exists $m_u \in M$ such that the performance q is optimal. This procedure defines a mapping D: $U \quad M$ such that for any $u \in U$, $D(u) = m_u$. The image of U under D, i.e. $M_D = D(U)$ will be called the optimal domain. M_D is the set of all $m \in M$ which are optimal with respect to at least one $u \in U$. (ii) To any $m \in M$ corresponds a subset $U_m \subset U$ such that the performance q is satisfactory in U_m and only there. In general, U_m is a proper subset of U. However, there might be some elements in M such that $U_m = U$, i.e. the performance for such m is satisfactory in the entire set U. Set of all such elements, M_S, (i.e. $M_S = \{m : G_m(u) <_q V(u)\}$) is called the satisfaction domain. (iii) Let $p(m,u) = G(m,u) - V(u)$ and let m^0 be such that $\min_M \max_U p(m,u) = \max_U p(m^0,u)$. m^0 is the element that has the "best worst" performance. Furthermore, for any pair $m,m' \in M$ introduce the difference $\Delta(m',m,u) = p(m',u) - p(m,u)$; $\Delta(m',m,u)$ can be also written in the form $\Delta_{m'}(m_1 u)$ if the explicit dependence on (m,u) is of interest. We can introduce then the set $M_A = \{m' : \max_U \min_M \Delta_m 0(m,u) \leq \max_U \min_M \Delta_{m'}(m,u)\}$.

Of particular interest are the sets M_D, M_S and the element m^0 which has the "best worst" performance. The following question seems to be of importance:

(i) Is m^0 always an element of M_S and if it is not, whether M_S is empty?

(ii) Is m^0 an element of M_D and if not, what is the relation between M_D and M_A?

(iii) What is the relation between the sets M_S and M_D? If $M_S \cap M_D$ is empty is M_S empty too?

Of special importance is the last question. From the example in the previous section, we already have a negative answer; i.e., there are classes of problems for which M_S is not empty although the intersection $M_S \cap M_D$ is void. In other words, there exist problems for which no satisfactory solution is optimal with respect to any disturbance. The question then can be rephrased: what is the class of problems for which $M_S \cap M_D$ is non-void whenever M_S is non-void? In other words, if there exists a satisfactory solution M_S then there exists an element of M (not necessarily M_S) which is optimal with respect to a given disturbance while still satisfactory in the entire U.

This question is of considerable practical importance. A number of optimization problems are constructive (at least potentially) in the sense that algorithm can be given to determine the optimal m (even if only approximately). If the intersection $M_S \cap M_D$ is empty then optimization will increase the global sensitivity beyond limits so that for any $m \in M_D$ U_m is a proper subset of U (without excluding the possibility of a non-empty M_S). If, on the other hand, $M_S \cap M_D$ is non-empty there exists at least one $m \in M_D$ such that $U_m = U$; the optimization procedure does not necessarily increase the sensitivity beyond the limits and the problem reduces to the selection of $u \in U$ such that $D(u) \in M_S \cap M_D$; i.e. the optimal m for that u is also an element of M_S. There exists then a chance (depending upon a proper selection of u) that the optimization will lead to a satisfactory performance.

In this paper we will have the space to present only some of the answers to these questions; in particular, those connected with the global sensitivity. A more fully discussion will be presented elsewhere.

GLOBAL SENSITIVITY AND OPTIMIZATION

In connection with the questions in the previous section, it should first be noted that $m^0 \in M_s$ whenever M_s is non-empty. However, m^0 is not necessarily a member of M_D. If

$m^0 \in M_D$ then $M_A = M_D$. On the other hand, if m^0 is not a member of M_D then M_D is only a proper subset of M_A.

Relation between the sets M_D, M_S and the global sensitivity for a broad class of important problems is given in the following theorem.

Theorem: Let M and U be compact and convex sets. Furthermore, let Δ_m, (m,u) be convex in m for each $u \in U$ and concave in u for each $m \in M$. Then if $m' \notin M_D$ there exists $m'' \in M_D$ such that the global sensitivity of $S_{m''}$ is smaller than the sensitivity of S_m.

Proof of this theorem is given in $^{(4)}$. The theorem gives the condition when a least sensitive (globaly) element m is always optimal with respect to some $u \in U$. If these conditions are satisfied one can do no better than by optimization (with respect to the global sensitivity). However, it should be noticed that the theorem does not state that $M_D \subset M_S$; i.e. that any optimal solution is sufficiently insensitive. The theorem only states that there are some such optimal solutions and the problem then reduces to the search for an element in U for which $D(u) \in M_S$. This of course is not necessarily a simple problem in itself but at least the existence of a satisfactory solution is ensured. One might also take a pragmatic approach and consider that a reasonable approximation to such an element in U will result in the behavior of the system which is sufficiently insensitive.

From the Theorem an interesting corollary has been derived stating that if there is no optimal solution $m \in M_D$, such that its global sensitivity is satisfactory (i.e. $M_D \cap M_S$ is void) then none of the elements from M is sufficiently insensitive (i.e. M_S is void).

As an illustration of the problems covered by Theorem, it has been shown that the linear dynamic system with a so-called quadratic performance functional satisfies the conditions of the theorem and therefore, the condition for the existence of a sufficiently insensitive element $m \in M$ reduces to the compactness and convexity of the sets U and M.

Finally, it should be noted that proof of the Theorem is based on the existence of the saddle point for

$\Delta_{m}'(m,u)$ but not on the existence of the game-theoretic solution for the original problem. Namely, the existence of the solution for the game in U and M is independent from the conditions in the Theorem. This could also be seen from the example considered in the first part of the paper.

REFERENCES

1. Mesarovic, M.D., "Decision-Making Under True Uncertainty", Systems Research Center report 10 (1962), Case Institute of Technology.
2. Macko, D.S., Mesarovic, M.D., "Uncertainties and Optimal Control Approach to Feedback Control Problems", Information & Control, (1965).
3. Mesarovic, M.D., "Satisfaction Approach To The Synthesis and Control of Systems", Third Allerton Conference on Systems and Circuit Theory, 1965.
4. Takahara, Y., "Multi-Level Systems and Uncertainties", Ph.D. Thesis, Case Institute of Technology.

ON THE COMPATIBILITY OF ADAPTIVE CONTROLLERS

S. P. Bingulac
"Boris Kidric" Institute of Nuclear Sciences
Belgrade, Yugoslavia

Abstract

In order to minimize the effects of plant parameter variations on system performance, the ultimate purpose of each plant adaptive system is to correlate the adjustments ΔC of controller parameters to the plant parameter variations ΔP, i.e. to define the vector relation

$$\Delta C = F(\Delta P) \ . \tag{a}$$

However, as parameters P vary, an important practical question arises; namely, to what extent the dependence (a) will maintain the desired system performance.

For an analysis of this problem the notion of compatibility of adaptive controller is introduced. The compatibility is used as a measure of the capability of selected adjustable controller parameters to compensate for the effects caused by the variations of plant parameters P.

As a first step, the notion and the index of compatibility in the small have been introduced, permitting the estimation of the controller compatibility in the vicinity of nominal or expected values of plant and controller parameters. An investigation in the small leads to the use of the first order sensitivity functions. It is also shown that the index of compatibility in the small may be determined using only the plant and controller sensitivity functions. This property considerably facilitates a quick selection of the best possible controller parameters to be adjusted by the adaptive system.

Introduction and Definitions

In the design of adaptive system [1], we are often faced with the problem of choosing the most appropriate structure of adaptive controller, as well as with the problem of selecting the best combination of available controller parameters to be adjusted by an adaptive loop. Suppose we are given a plant, characterized by its variable parameters

$$P = \{p_m\} \in D_p \ , \qquad (m = 1,2,\ldots M) \tag{1}$$

where D_p is a domain of expected plant parameter variations. Also, we have chosen the controller with N^* free parameters

$$C^* = \{c_n^*\} \qquad (n = 1,2,\ldots N^*) \ . \tag{2}$$

The question is now which and how many controller parameters (Fig. 1)

$$C = \{c_n\} \qquad (n = 1,2,\ldots N) \ N < N^* \tag{3}$$

are to be adjusted by adaptive system in order to perform satisfactory minimization of the given performance index. As a performance index we will take an average or instantaneous error criterion

$$J(P,C) = \overline{e(t,P,C)^2} \quad \text{or} \quad J(P,C) = e(t,P,C)^2 \tag{4}$$

where $e(t,P,C)$ is an error function depending upon the system output $y(t)$ and parameters P and C.

As the number N of adjustable controller parameters strongly influences the complexity of adaptive system, it is desired to keep N as small as possible but without sacrificing the performance of the whole system. Attempting to obtain one-to-one correspondence between variable plant and adjustable controller parameters [2], [3] is not applicable in the general case, and it also leads to systems where N is always equal to number M of variable plant parameters.

To investigate these questions, the notion of compatibility of adaptive controller to the given plant has been introduced recently [1]. The compatibility of the controller represents the capability of the adjustable controller parameters (3) to compensate for the undesired effects caused by the variations of plant parameters (1). The controller compatibility may be ascertained by applying one of the existing self-optimizing methods [4], [5], [6], [7], for various points in the P-space of the variable plant parameters. From the practical point of view however, it would be desirable to have a procedure for quicker determination of controller compatibility. Unfortunately, the dependence of system performance on parameters is always higher nonlinear, and therefore, an analytical investigation of the compatibility in the large is rather complex problem.

As the first step in analyzing the controller compatibility, this paper introduces the notion of complete compatibility in the small and defines the index of compatibility, permitting the estimation of controller compatibility in the vicinity of nominal or expected values of plant and controller parameters. An investigation in the small leads to the use of the first order sensitivity functions.

It will also be shown that the index of compatibility in the small may be determined using only the plant and controller sensitivity functions defined as

$$\begin{aligned} U(t) &= \{u_m(t)\} = \{\partial y(t)/\partial p_m\} \\ V(t) &= \{v_n(t)\} = \{\partial y(t)/\partial c_n\} , \end{aligned} \tag{5}$$

$y(t)$ being system output.

This property considerably facilitates quick selection of the best possible controller parameters to be adjusted by the adaptive system. Let us suppose further that in the design stage of control system the optimum values which minimize the performance index (4) in the case when plant parameter have the nominal or expected values

$$P = P^o \tag{6}$$

$$C = C^o \tag{7}$$

are determined. Therefore, C^o corresponds to

$$J_{min} = \min_C \left[J(C,P^o)\right] .$$

In this case error function becomes

$$e^{o}(t) = e(t,P^{o},C^{o}) \ . \tag{8}$$

Compatibility in the Large

Now we may state the definition of the compatibility more precisely. Let the adaptive system adjust the controller parameters so that they follow the relation

$$C = F(P) \ . \tag{9}$$

If this relation satisfies the condition

$$J[P^{o},C^{o}] = J[P,F(P)] \ , \tag{10}$$

the controller will be referred to as completely compatible in the large.

Instead of (10) we may consider the condition

$$e^{o}(t) = e[t,P,F(P)] \tag{11}$$

which is obviously a sufficient condition for condition (10).

Unfortunately, in practice and particularly in the case of multi-loop nonlinear systems the condition (10) of complete compatibility in the large is not likely to be satisfied. As has been said previously, due to inherently nonlinear dependence of the performance index on parameters, the analytical investigation of the compatibility in the large is a rather complex problem.

As the first step in the investigation of compatibility, one may try to estimate the controller compatibility in the small, i.e. in the vicinity of nominal plant and optimum controller per meter values

$$P^{o} \text{ and } C^{o} \ . \tag{12}$$

Moreover, compatibility in the small may be considered as a necessary (but not sufficient) condition for the satisfactory operation of the adaptive system in the whole domain D_p.

Compatibility in the Small

For small variations of plant and controller parameters

$$\|\Delta P\| \ll \|P^{o}\| \quad \text{and} \quad \|\Delta C\| \ll \|C^{o}\| \ ,$$

the minimized error function e(t) may be written as

$$e(t) = e^{o}(t) + \Delta e(t) \ . \tag{13}$$

Thus, variation $\Delta e(t)$ of the error function may be expressed [8] as a sum of two terms,

$$\Delta e(t) = e_p(t) + e_c(t) ,$$

$$e_p(t) = \sum_{m=1}^{m=M} \partial y(t)/\partial p_m \, \Delta p_m \tag{14}$$

$$e_c(t) = \sum_{n=1}^{n=N} \partial y(t)/\partial c_n \, \Delta c_n$$

where $e_p(t)$ and $e_c(t)$ originated from the variations of P and C, respectively.

Then, using (5), the variation $\Delta e(t)$ becomes

$$\Delta e(t) = U(t)'\Delta P + V(t)'\Delta C , \tag{15}$$

where symbol (') denotes vector or matrix transposition.

According to the previously given definitions (10) and (11), for the condition of complete compatibility in the small, one obtains

$$\Delta e(t) = U(t)'\Delta P + V(t)'\Delta C = 0 \tag{16}$$

which should be satisfied for every t. On the other hand, relation (9) implemented by the adaptive system, in the case of small parameter variations, becomes

$$\Delta C = A\Delta P \tag{17}$$

where

$$A = \{a_{nm}\} = \{\partial f_n/\partial p_m\}$$

is (NxM) constant matrix.

Introducing (17) into (16), for the condition of complete compatibility in the small, one obtains

$$U(t)' = - V(t)'A . \tag{18}$$

Thus, the controller (3) is compatible in the small to the given plant (1) if there exists a linear, constant coefficient, transformation A transforming vector V(t) into U(t). From (18) matrix A is given by

$$A = - [V(t_s)\ V(t_s)']^{-1}\ [V(t_s)\ U(t_s)'] \tag{19}$$

where $t = t_s$ is any instant in the time interval in which the behavior of the system is considered. It is more practical however, to calculate matrix A from the average values

$$A = - [\ \overline{V(t)\ V(t)'}\ ]^{-1}\ [\ \overline{V(t)\ U(t)'}\ ] \tag{20}$$

which evidently in the case of (18) gives the same results as (19).

Index of Compatibility in the Small

From the above analysis it may be concluded that the sensitivity functions (5) of the unperturbed system (12) may serve as an indication whether the chosen controller is compatible to the given plant. Hence, as the condition of complete compatibility in the small is given by (18), an index of compatibility may be derived, indicating how much the components of the M dimensional vectors U(t)' and V(t)'A differ from each other. To this end, consider the general case in which the condition (18) is not satisfied. Introducing a vector E(t) as

$$E(t)' = U(t)' + V(t)'A \tag{21}$$

where matrix A is given by (20), an index of compatibility in the small may be defined by

$$k = \frac{\overline{E(t)'E(t)}}{\overline{U(t)'U(t)}} \quad . \tag{22}$$

From (20) and (21) it follows that the scalar quantity k satisfies

$$0 \leq k \leq 1 \ . \tag{23}$$

As the index of compatibility is smaller, the controller is more compatible to the given plant. Therefore, the value of k may be useful in quick estimation of the degree of controller compatibility. In other words, from the value of k one may easily conclude to what extent the adjustment of vector C will, in the vicinity of nominal parameter values (12), compensate the error $e_p(t)$, given by (14).

Investigating and comparing the indices of compatibility for different controller structures, and taking as adjustable only several c_n $(n = 1,2,\ldots N)$; $N < N^x$, for all N^x controller parameters, available for adjustment, it is possible to select the optimum controller structure and the best possible combination of adjustable controller parameters. In the case of linear control systems, application of the method of sensitivity points [9], [10] further simplifies the determination of the index of compatibility, because all required sensitivity functions (5) may be simultaneously generated on the computer using only two identical models of the system.

Examples

In order to illustrate the procedure of analyzing the controller compatibility let us consider the linear three loop system [1], shown in Fig. 2. The nominal values of variable plant parameters are

$$P^o = \begin{Bmatrix} p_1^o \\ p_2^o \\ p_3^o \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix}$$

while optimal values of controller parameters, determined in the usual manner [4] are

$$C^o = \begin{Bmatrix} c_1^o \\ c_2^o \\ c_3^o \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix}$$

As an error function e(t) we will take the deviation of step response of the output variable y(t) from $y^o(t)$, i.e.

$$e(t) = y(t) - y^o(t) \, ,$$

where $y^o(t)$ corresponds to the nominal parameter values (12).

Fig. 3 and 4 show $y^o(t)$ together with the sensitivity functions $u_m(t)$ and $v_n(t)$, respectively.

We will separately study two following cases:

a) two plant parameters (p_1,p_2) are variable, and

b) three plant parameters (p_1,p_2,p_3) are variable.

Case a)

In the case a), if both c_1 and c_2 are allowed to be adjustable, index of compatibility k becomes zero. Therefore, in this case the controller (c_1,c_2 adjustable) is completely compatible in the small to the given plant (p_1,p_2 are variable). However, if we are restricted to one-parameter adaptive system, the question is which parameter among

$$c_1, \ c_2 \text{ and } c_3$$

is the most appropriate for adjustment.

The computer results give

1. $k_1 = 0.74$ if c_1 is adjustable
2. $k_2 = 0.33$ if c_2 is adjustable
3. $k_3 = 0.81$ if c_3 is adjustable .

Therefore, if p_1 and p_2 are variable it is more advantageous to adjust parameter c_2, by one-parameter adaptive system, then c_1 or c_3.

Experiments on the computer are in accordance with the obtained conclusion.

Case b) - Three variable plant parameters (p_1, p_2, p_3)

It may be shown [11], that in the case b) index k is zero if we have three-parameter adaptive system (c_1,c_2 and c_3 are adjustable). As the index of compatibility is zero, this case corresponds to the completely compatible controller.

Similarly, we are interested in the question which one of the three possible two-parameter adaptive systems

$$((c_1,c_2), \ (c_1,c_3) \ \text{ or } \ (c_2,c_3) \text{ are adjustable})$$

is most convenient. Applying the similar procedure, as in the case a) we obtain

1. $k_1 = 0.437$ if (c_1, c_2) are adjustable
2. $k_2 = 0.326$ if (c_1, c_3) are adjustable
3. $k_3 = 0.099$ if (c_2, c_3) are adjustable .

Evidently the most convenient two-parameter adaptive system is that one which adjusts controller parameters c_2 and c_3. Recent computer simulations [1], [12] have experimentally proved the conclusion that it is most convenient to adjust parameters c_2 and c_3.

In order to select the best controller structures, different structures may be analyzed similarly, and their indices of compatibility are to be compared with those just calculated.

Discussion and Conclusion

In this paper it is shown that the index of compatibility in the small may be investigated considering only the plant and controller sensitivity functions. According to the examples it may be concluded that calculation and comparison of indices of compatibility for different controller structures and various combinations of adjustable controller parameters permits and facilitates a quick selection of the best possible controller parameters to be adjusted by the adaptive system. On the other hand, in the general case, for the estimation of the compatibility in the large the repeated self-optimization is the only possibility. However, it may be shown that for systems of the type studied here [1] (Fig. 2), the condition $k = 0$ also guarantees complete compatibility in the large. Hence, the controller (c_1,c_2 and c_3 are adjustable) will be completely compatible in the large to the given plant (p_1, p_2 and p_3 are variable).

Acknowledgement

The author wishes to thank Prof. P. V. Kokotović, for his very valuable discussions and suggestions.

References

[1] Kokotovic, P. V., J. Medanic, M. Vuskovic, S. P. Bingulac, Sensitivity Methods in the Experimental Design of Adaptive Control Systems, III-rd IFAC Congress, London, 1966.

[2] Puri, N. N., C. N. Weygandt, Multivariable Adaptive Control Systems, Proc. Joint Automatic Control Conf., New York, 1962.

[3] Whitaker, H. P., Design Capabilities of Model Reference Adaptive Systems, Proc. NEC, Chicago, 1962.

[4] Bingulac, S. P., P. V. Kokotovic, Automatic Optimization of Linear Control Systems on Analog Computers, Proc. of the International Association on Analog Computation, (IACA), Vol. VII, No. 1, Jan. 1965, Bruxelles.

[5] McBride, L. E., K. Steiglitz, Iterative Methods for System Identification, Technical Report No. 15, June 1966, Communication Lab., Princeton University, N. J.

[6] Kokotovic, P. V., S. P. Bingulac, J. Medanic, Some Approaches to the Reduction of Control System Sensitivity, Proc. of 3rd Allerton Conf. on Circuit and System Theory, 1965.

[7] Meissinger, H. F., Parameter Optimization by an Automatic Open Loop Computing Method, IV-th AICA Congress, Brighton, 1964.

[8] Bingulac, S. P., Minimum Sensitivity Adaptive Systems, Proc. of an International IFAC/ETAN Symposium on Sensitivity Method, Editor Lj. Radanovic, Pergamon Press, 1966.

[9] Kokotovic, P. V., Simultaneous Computation of Sensitivity Functions, Pupin Control Lab. Report, July 1962; also published in R. Tomovic, Sensitivity Analysis of Dynamic Systems, pp. 57-78, McGraw Hill, New York, 1965.

[10] Kokotovic, P. V., Sensitivity Points Method in Analysis and Optimization of Linear Control Systems, Automatic and Remote Control, No, 12, Dec. 1964, Moscow.

[11] Bingulac, S. P., Investigation of the Compatibility in Adaptive Control Systems, IBK Publication No. 450b, Belgrade, June 1966.

[12] Bingulac, S. P., A Transformation Improving the Stability of Adaptive Systems, Fourth Allerton Conf. 1966, University of Illinois, Urbana, Illinois.

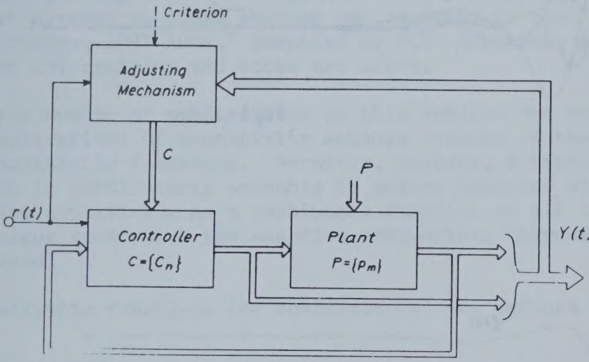


Fig 1

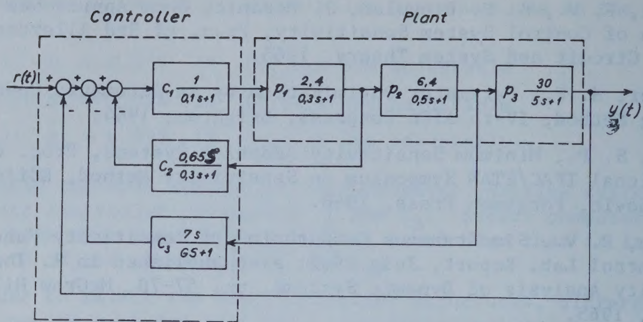

Controller
Plant
r(t)
+
+
+
-
-
-
$C_1 \frac{1}{0,1s+1}$
$p_1 \frac{2,4}{0,3s+1}$
$p_2 \frac{6,4}{0,5s+1}$
$p_3 \frac{30}{5s+1}$
y(t)
$C_2 \frac{0,65s}{0,3s+1}$
$C_3 \frac{7s}{6s+1}$

Fig. 2

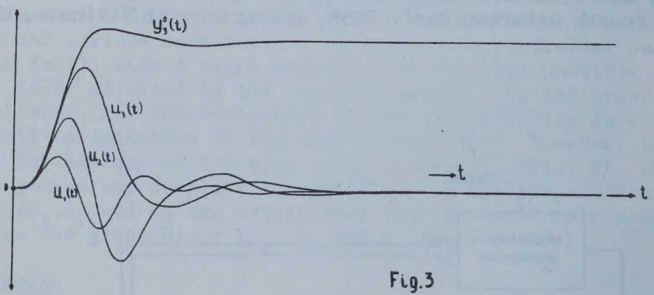


Fig.3

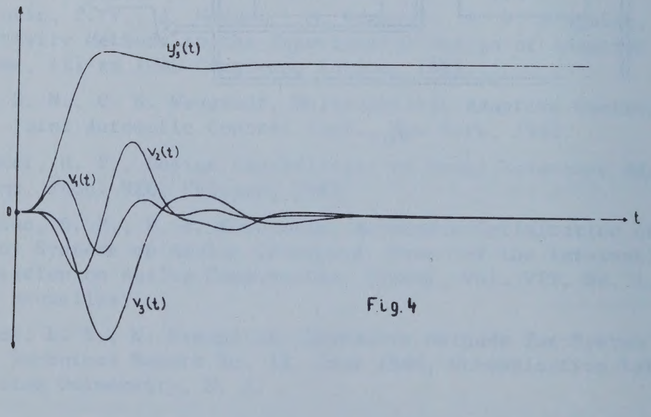


Fig. 4

THE APPLICATION OF SENSITIVITY METHODS TO THE CONTROL OF ADAPTIVE SYSTEMS

AGOSTINO BURZIO
University of Santa Clara
Santa Clara, California
ESL, Inc., Palo Alto, California

RICHARD C. DORF
University of Santa Clara
Santa Clara, California

ABSTRACT

This paper discusses a class of adaptive systems in which sensitivity coefficients are used as a direct control over the adaptive process. The adaptive system compares the time response of the controlled system with the desired or ideal response. Then sensitivity coefficients and logic circuits are used to control the variation of certain parameters of the controlled system in such a way as to decrease the difference between the two responses. Neither test signals nor iterative procedures are needed, and adaption is rapid. Included is a technique which does not require system identification prior to, or during adaption. Excellent results have been obtained in the analog computer simulation of several systems; examples are given.

INTRODUCTION

In recent years, studies of the sensitivity of system response to parameter variations have shown its importance in providing a better understanding of the behavior of systems under varying conditions [1]. For example, the problems of optimization of performance and plant identification have been approached through sensitivity analysis. An indication of the effort devoted to these studies can be found in the "Bibliography on Sensitivity Theory, 1937-1964," compiled by P.V. Kokotovic and R.S. Rutman [4], in which 157 articles and books are cited.

The large number of publications on this subject has not prompted many practical applications of sensitivity methods because of the difficulty in obtaining sensitivity functions. Recently, however, a technique has been devised which is particularly amenable to analog computer simulation and which provides sensitivity as a continuous function at all times [2]. This is the technique upon which the adaptive controllers introduced in this paper are based.

The sensitivity functions (or coefficients) are defined in the time domain as

$$\underline{v}_i(t) \triangleq \frac{\partial \underline{x}}{\partial p_i} , \tag{1}$$

where p_i is a parameter of the system which is excited by inputs $\underline{r}(t)$. The system is defined by the relation

$$\underline{F}(\dot{\underline{x}}(t),\ \underline{x}(t),\ \underline{r}(t),\ \underline{p}(t),\ t) = 0 . \tag{2}$$

Some functions have been used successfully in parameter optimization studies on the analog computer [2,3]; and although linear systems have been used as examples, the same procedures can be applied directly to nonlinear systems. The optimization methods, which are based on an iterative procedure, may also be used to control adaptive processes; however, perturbation signals have to be introduced into the system.

It is possible, using these same sensitivity functions, to devise certain system philosophies which provide adaption without the need of perturbation or dither signals. The exact mathematical description of the behavior of these adaptive controllers has not been attempted; however, some very salutary results have been obtained through qualitative reasoning and analog computer simulation. Extensive experimentation has also shown that systems with adaptive capabilities do not have to be optimally designed to give an excellent performance.

Before continuing this discussion, the definition of an adaptive system should be given. Even though no universally accepted definition exists, the following is proposed: An adaptive system is one which continually monitors its performance and adjusts its parameters or structure in order to maintain the prescribed performance.

The problem which will be considered for the illustration of the sensitivity methods of adaption is as follows. Given a system whose response under different environmental conditions varies considerably, one wishes to design a controller which will cause the system to have the same desired response under all possible conditions.

It has been shown [5] that it is possible, in some cases, to design a conventional (non-adaptive) controller that will result in a system which will perform as well as an adaptive system. It must be assumed, however, that the transfer function of the system is known. Most adaptive schemes also require that system identification be carried out before or during the adaptive process.

Here, it will first be assumed that the equations describing the system are known, and then a scheme which does not require plant identification will be discussed. For simplicity, a single parameter and a single system output will be chosen for control; however, the results can readily be extended to the multi-parameter and multiple output case.

The Basic Sensitivity Equations.

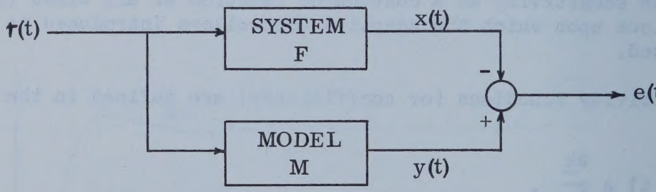


Figure 1.

Consider the system F (Figure 1) which is required to have a given response y(t) to an input r(t), and let M represent a "Model" system whose output is y(t) when excited by r(t) . Also let the error function e(t) be defined as the difference between y(t) and the actual system response, x(t), which is to be controlled by a system parameter p . Then the time-derivative of e(t) can be written as

$$\frac{de}{dt} = \frac{\partial e}{\partial p} \cdot \frac{dp}{dt} + \frac{\partial e}{\partial t} . \qquad (3)$$

Since the variational effect of p on e is

$$\frac{\partial e}{\partial p} = \frac{\partial y}{\partial p} - \frac{\partial x}{\partial p} , \tag{4}$$

and since y is unaffected by changes in p,

$$\frac{\partial e}{\partial p} = - \frac{\partial x}{\partial p} = -v . \tag{5}$$

The change of p with time can be expressed in terms of the error variations by rearranging equation (3):

$$\frac{dp}{dt} = \frac{1}{v} \left(\frac{\partial e}{\partial t} - \frac{de}{dt} \right), \tag{6}$$

where $\frac{\partial e}{\partial t}$ represents the time rate of change of error with the parameter p considered to be constant. Equation (6) describes the effect of the control parameter on the error; and these variables are related through the sensitivity function v . Various adaption schemes can be devised using this relationship.

The Ideal Adaptive System.

Let it be required that $\frac{de}{dt}$ be kept equal to zero (assuming that the initial error is zero, this would mean perfect tracking of y by x). Then, from equation (6), the variation of p must be

$$\frac{dp}{dt} = \frac{1}{v} \cdot \frac{\partial e}{\partial t} . \tag{7}$$

The generation of the sensitivity function has been discussed (2); now, to obtain $\frac{\partial e}{\partial t}$ as a continuous function, it is necessary to provide a separate simulation, as follows.

Let the system F of Figure 1 be described by the characteristic equation

$$F(\dot{\underline{x}}, \underline{x} , p, t) = r(t) . \tag{8}$$

Taking the partial derivative with respect to t, while holding p constant, one obtains an equation of the form

$$G(\dot{\underline{x}}', \underline{x}', p, t) = r'(t), \tag{9}$$

where $\underline{x}'$ and r'(t) are defined as

$$r'(t) \triangleq \frac{\partial r(t)}{\partial t} = \frac{dr(t)}{dt} , \quad \underline{x}' \triangleq \frac{\partial \underline{x}}{\partial t} \text{ (p = constant)}. \tag{10}$$

Similarly, if the model dynamics are described by

$$M(\dot{\underline{y}}, \underline{y}, t) = r(t), \tag{11}$$

differentiation with respect to t yields

$$Q(\dot{\underline{y}}', \underline{y}', t) = r'(t), \tag{12}$$

with the definition

$$\underline{y}' \triangleq \frac{\partial \underline{y}}{\partial t} = \frac{d\underline{y}}{dt} \qquad (p = \text{constant}). \tag{13}$$

Thus the desired function,

$$\frac{\partial e}{\partial t} = \frac{\partial y}{\partial t} - \frac{\partial x}{\partial t} = y' - x', \tag{14}$$

can be obtained by simulating the systems G and Q and using as the input the derivative of the original forcing function. The complete adaption scheme would take the form of Figure 2.

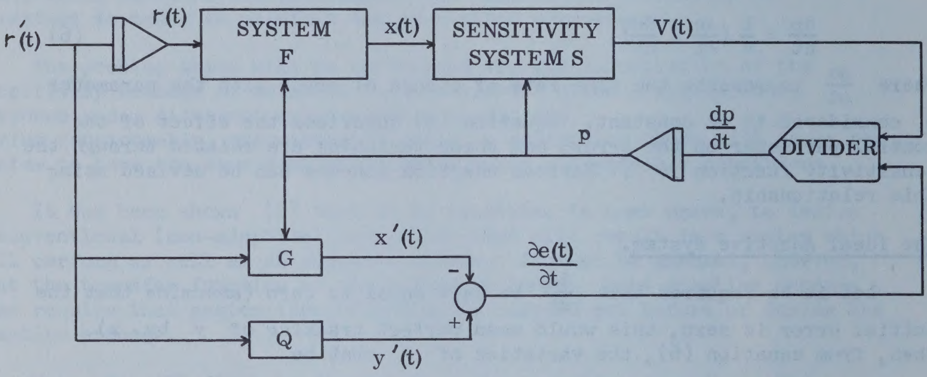


Figure 2

Although in theory the system of Figure 2 gives ideal results (i.e., zero error at all times), its practical use is limited by various factors:

1) The dynamics of the system to be controlled must be known;

2) The parameter p must be available for control;

3) The derivative of the forcing function of the system must be available;

4) For complex systems the simulation will also be complex, since the system's dynamics, with slight variations, must be re-simulated twice (G and S in Figure 2);

5) Due to time delays and disturbances in the simulation, the error will shift from zero; and, since only the derivative of the error is monitored for control, the accumulated error signal could eventually become very large.

Because of these severe limitations, the adaption scheme of Figure 2 was not utilized. Instead, a slight change in the adaption philosophy provided a more practical solution of the problem.

Bounded-Error Adaptive System. Consider the following criterion: if the magnitude of the error (see Figure 1) is increasing, let an adaption signal be generated to counteract this effect and reduce the error; but when the error is decreasing in magnitude no adaption should be attempted. A reasonable amount of adaption, where needed, could be provided by making the adaption signal $\frac{dp}{dt}$ such that it tends to reverse the error velocity. Thus, based on equation (6), let

$$\frac{dp}{dt} = \frac{2}{v} \cdot \frac{\partial e}{\partial t} = \frac{2}{v} \frac{de}{dt} \quad , \tag{15}$$

since, just prior to the adaption transient, $\frac{dp}{dt} = 0$ and therefore

$$\frac{\partial e}{\partial t} = \frac{de}{dt} \quad .$$

This criterion for the adaptive controller can be described mathematically:

$$\frac{dp}{dt} = \begin{cases} 0 & \text{when } \operatorname{sgn} \frac{de}{dt} = - \operatorname{sgn} e \\ \frac{2}{v} \cdot \frac{de}{dt} & \text{when } \operatorname{sgn} \frac{de}{dt} = \operatorname{sgn} e \quad . \end{cases} \tag{16}$$

In order to eliminate steady-state errors which may arise due to the fact that $\frac{de}{dt}$ can be zero even though e is non-zero, one can modify equation (16) as follows:

$$\frac{dp}{dt} = \begin{cases} 0 & \text{when } \operatorname{sgn} \frac{de}{dt} = -\operatorname{sgn} e \\ \frac{2}{v} (\frac{de}{dt} + ke) & \text{when } \operatorname{sgn} \frac{de}{dt} = \operatorname{sgn} e \quad . \end{cases} \tag{17}$$

Here k is some positive constant.

Practical implementation of this adaptive scheme imposes other restrictions on equation (17). For example, the value of v cannot be allowed to go to zero; and, since in many problems v can have both positive and negative values, only the absolute magnitude of v must be used in determining the correction signal dp/dt, thus necessitating the use of logic circuits in the adaptive loop. Moreover, when the magnitude of the sensitivity function is small, no adaption signal should be generated since, in this case, even a large variation in p would produce a very small change in the system response. Thus equation (17) should be modified further to read

$$\frac{dp}{dt} = \begin{cases} \frac{2}{v} (\frac{de}{dt} + ke) & \text{when } \operatorname{sgn} \frac{de}{dt} = \operatorname{sgn} e, \ v > V_{min} \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

V_{min} being a small positive number chosen as the lowest value of v for which adaption is still desirable.

The complete scheme is shown in Figure 3.

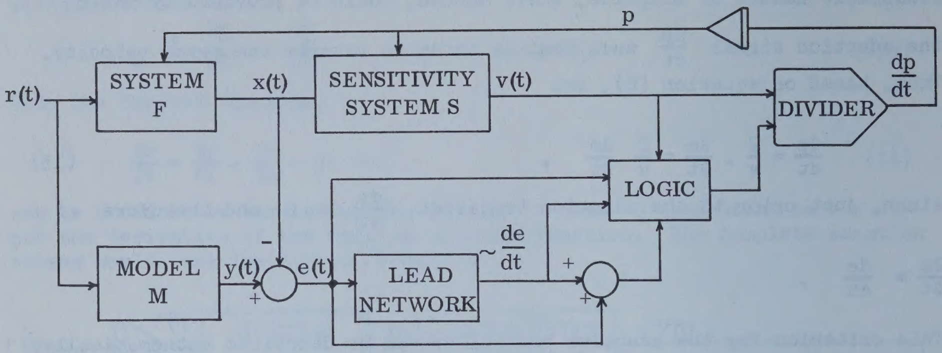


Figure 3

It should be pointed out that this adaptive system still has two of the disadvantages of the previous system; namely,

1) The dynamics of the system to be controlled must be known;

2) The parameter p must be available for control.

However, there are many systems which satisfy the above conditions; and the example below demonstrates the feasibility of such an approach.

Example. Consider the third-order system described by

$$\dddot{x} + a\ddot{x} + b\dot{x} + x = r(t),$$

and let b be the controllable parameter and the parameter a be subject to wide variations. It is required that, at all times, the system response approximate as closely as possible that of the second-order system

$$\ddot{y} + \dot{y} + y = r(t),$$

regardless of the value of a .

The sensitivity of the response to variations in b can be found by simulating the sensitivity equation

$$\dddot{v}_b + a\ddot{v}_b + b\dot{v}_b + v_b = -\dot{x},$$

v_b being defined as

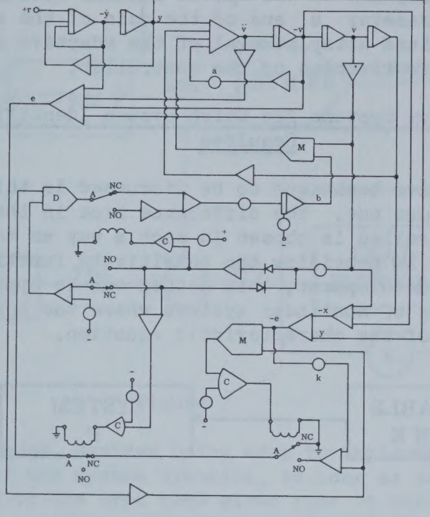
$$v_b \triangleq \frac{\partial x}{\partial b} .$$

The other two functions which are required as inputs to the adaptive controller, namely e and $\dot{e}$ of Figure 3, can be generated with simple summation amplifiers, since

$$e(t) = y(t) - x(t)$$

and

$$\dot{e}(t) = \dot{y}(t) - \dot{x}(t) .$$



+r
-y
y
v̇
-v
v
a
e
M
D
A
NC
NO
b
C
-x
-e
y
k

Figure 4.

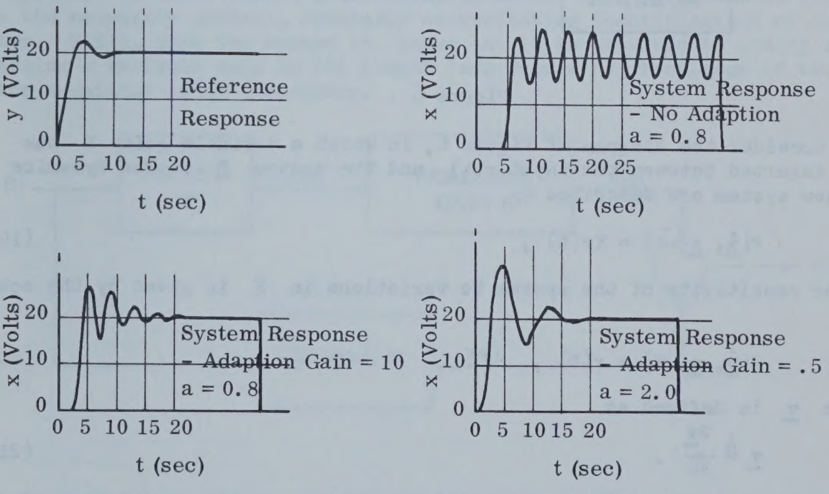


Figure 5.

The logic functions required by equation (18) are performed by means of diodes, comparators, and relays.

The complete analog computer diagram of the adaptive system simulation is shown in Figure 4; and the results obtained from this simulation appear in Figure 5. These results show that, using the adaptive philosophy of Equation (18), the response of the system remains satisfactory for large variations of the parameter a and of the gain of the adaptive loop. The system response obtained after removal of the adaptive controller gives an indication of the effectiveness of the controller.

Bounded-Error Adaptive Systems for which System Identification is not Required

The third adaptive technique to be discussed in this paper is basically similar to the previous one. The difference lies in the fact that the parameter to be controlled is chosen in such a way as to make system identification unnecessary in providing the sensitivity function. As will become apparent during its development, this method can be applied to linear systems and to the types of nonlinear systems where the nonlinearity appears in the coefficients of the characteristic equation.

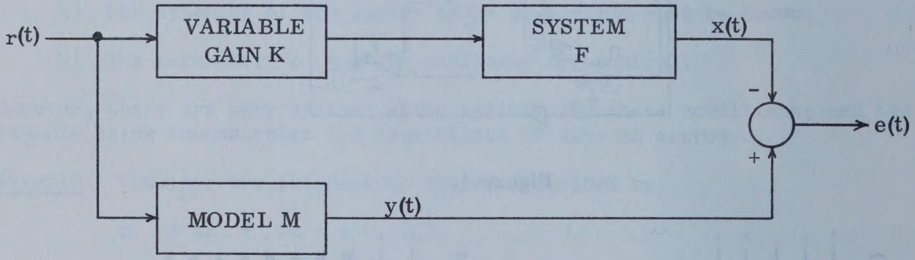


Figure 6

Consider the diagram of Figure 6, in which a variable gain K has been inserted between the input r(t) and the system F . The dynamics of the new system are described by

$$F(\dot{\underline{x}}, \underline{x}, t) = Kr(t) , \qquad (19)$$

and the sensitivity of the system to variations in K is given by the equation

$$F(\dot{\underline{v}}, \underline{v}, t) = r(t) , \qquad (20)$$

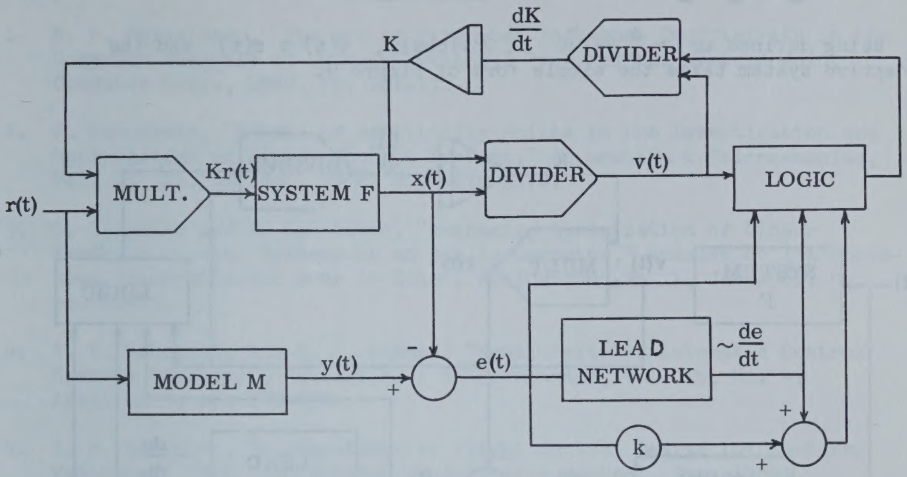
where $\underline{v}$ is defined as

$$\underline{v} \triangleq \frac{\partial x}{\partial K} . \qquad (21)$$

Obviously, equation (20) is not valid if F is a nonlinear function of $\underline{x}$ and $\dot{\underline{x}}$; and, for such systems, this adaption scheme cannot be used. For all other types of systems, equations (20) follows directly from (19); and a comparison of these two equations shows that the sensitivity function is simply

$$\underline{v} = \underline{x}/K . \qquad (22)$$

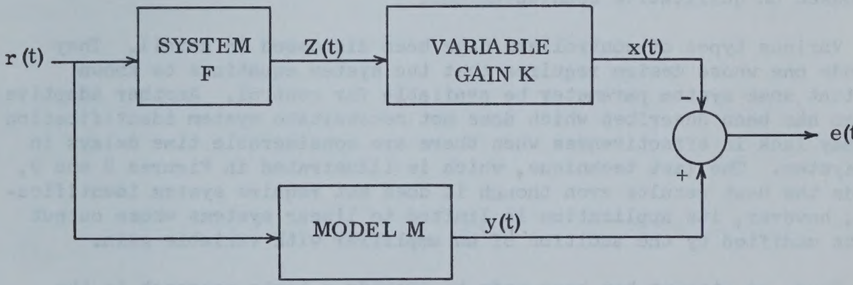
Again, if only one output of the system, x(t), is considered, the adaptive system takes the form of Figure 7.



dK/dt
K
DIVIDER
Kr(t)
MULT.
SYSTEM F
x(t)
DIVIDER
v(t)
LOGIC
r(t)
-
+
y(t)
e(t)
MODEL M
LEAD NETWORK
~ de/dt
+
k
+

Figure 7

This adaption scheme, besides being easy to implement, does not require any knowledge of the system dynamics, as long as any nonlinearity of the system is limited to the type that gives rise to nonlinear coefficients in the characteristic equation. However, the introduction of a variable gain which modulates the input to the system (so as to provide adaptive control) can sometimes lead to instability because of time delays inherent in the system. In these cases, a different parameter must be chosen to provide the necessary control, seemingly necessitating identification of the system. Still, when the system is linear and it is possible to modify it by adding a variable gain to its output (see Figure 8), knowledge of the system equations is not necessary.



r (t)
SYSTEM F
Z(t)
VARIABLE GAIN K
x(t)
-
+
e(t)
MODEL M
y(t)

Figure 8

Referring to Figure 8, if the system is linear and is described by

$$\dot{\underline{x}} + A\underline{x} = \underline{d}Kr(t), \qquad (23)$$

the sensitivity equation is simply

$$\dot{\underline{v}} + A\underline{v} = \underline{d}r(t), \tag{24}$$

$\underline{v}$ being defined as $\underline{v} \overset{\Delta}{=} \partial\underline{x}/\partial K$. Obviously, $v(t) \equiv z(t)$ and the adaptive system takes the simple form of Figure 9.

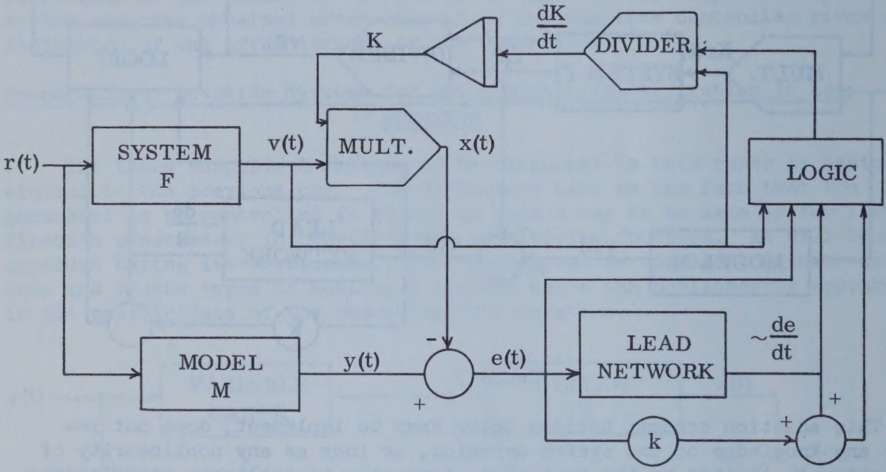


Figure 9.

CONCLUSIONS

This paper has presented a brief study of the application of sensitivity to the design of adaptive controllers. No attempt has been made to describe their exact behavior analytically; rather, the designs are based on qualitative considerations.

Various types of controllers have been discussed in detail. They include one whose design requires that the system equations be known and that some system parameter be available for control. Another adaptive system has been described which does not necessitate system identification but may lack in effectiveness when there are considerable time delays in the system. The last technique, which is illustrated in Figures 8 and 9, yields the best results even though it does not require system identification; however, its application is limited to linear systems whose output can be modified by the addition of an amplifier with variable gain.

Thus, an attempt has been made to provide a basic approach to the design of adaptive systems where sensitivity coefficients are used for control. Computer simulations of these adaptive controllers have been successful, showing that the application of these techniques to actual systems is feasible.

BIBLIOGRAPHY

1. H. F. Meissinger, "The Use of Parameter Influence Coefficients in the Computer Analysis of Dynamic Systems," Proc. of the Western Joint Computer Conf., 1960, pp. 181-192.

2. P. Kokotovic, "Method of Sensitivity Points in the Investigation and Optimization of Linear Control Systems," Automatika i Telemekhanika, Vol. 25, No. 12, Dec. 1964; pp. 1670-1676.

3. S. Bingulac and P. Kokotovic, "Automatic Optimization of Linear Feedback Control Systems on an Analog Computer," Annales de l'Association Internationale pour le Calcul Analogique, No. 1, Jan 1965; pp. 12-17.

4. P. V. Kokotovic and R. S. Rutman, "Sensitivity of Automatic Control Systems (Survey)", Automatika i Telemekhanika, Vol. 26, No. 4, April 1965; pp. 730-750.

5. I. M. Horowitz, "Linear-Adaptive Flight Control Design for Re-Entry Vehicles," IEEE Transactions on Automatic Control, Jan. 1964; pp. 90-97.

A TRANSFORMATION IMPROVING STABILITY OF ADAPTIVE SYSTEMS

S. P. Bingulac
"Boris Kidric" Institute of Nuclear Sciences
Belgrade, Yugoslavia

Introduction

One of the common practices in designing model reference adaptive systems [1], [2], [3], [4] consists in adjustment of controller parameters

$$C = \{c_n\} \; ; \; (n = 1,2,\dots,N) \qquad (1)$$

according to the gradient algorithm

$$\dot{C} = -h \; \mathrm{grad}_c [J(C)]; \quad (\dot{\cdot}) \equiv \frac{d}{dt} \; . \qquad (2)$$

The stability problems of this type of adaptive systems are usually analyzed for the case of small and/or slow (iterative) parameter variations [3], [5], [6], [7], [8]. In generalizing to more realistic situations, it is often intuitively assumed that the assured stability for slow and small parameter variations will considerably increase the chances that the system will be also stable for high speed and large parameter variations. To complete the analysis, this assumption always is checked experimentally by computer simulation. Computer experience, [1], [3], [5], has shown that complexity and importance of stability problem become pertinent as the dimensionality N of adjustable vector (1) is increased.

In solving the problems of parameter optimization of system identification, where iterative adjustment of vector C is performed, it is recognized also that the iterative version of the gradient algorithm (2), i.e.

$$\Delta C^{(i+1)} = -h \; \mathrm{grad}_c \; J[C^{(i)}] \; , \qquad (3)$$

cannot generally guarantee the optimization process which will be both fast and convergent. To cope with these situations, it has been shown recently [9], [10], [11] that the iterative algorithm

$$\Delta C^{(i+1)} = -H \; \mathrm{grad}_c \; J[C^{(i)}] \; , \qquad (4)$$

obtained from (3) substituting scalar h by nonsingular matrix $H = \{h_{nm}\}$, where

$$H = \{h_{nm}(c)\} = \{\partial^2 J(C)/\partial c_n \partial c_m\}^{-1} \; ,$$

$$(n,m = 1,2,\dots N) \qquad (5)$$

is much more appropriate for larger dimensionality N of adjustable vector C. However, as elements h_{nm} depend upon the system parameters it is necessary to recalculate matrix H for almost every iteration. This, of course, requires additional computer time in digital computation, while in analog computer application necessitates additional computer components.

Synthesis of Multiparameter Adaptive Systems

Following the usual practice in designing adaptive systems, it is reasonable, in the multiparameter case, to investigate the possibilities of application of the algorithm (4) which in the case of continuous adaptation becomes

$$\dot{C} = -H \operatorname{grad}_c [J(P,C)] \; . \qquad (6)$$
$$P = \{p_m\}; \; (m = 1,2,\ldots M)$$

Vector P denotes variable plant parameters, whose effects on system dynamics have to be compensated for by the adjustment of controller parameters C.

The dependence

$$h_{nm} = h_{nm}(P,C)$$

which generally holds, requires on-line calculation of the matrix H and its continuous readjustment in the adapting mechanism. As this requirement considerably complicates the practical implementation of the adaptive system, based on (6), it would be highly desirable to investigate the situations in which on-line recalculation and readjustment of H is not necessary.

Determination of the Matrix H

In the case of minimizing an average or instantaneous error criterion

$$J(P,C) = \overline{e(t)^2} \qquad \text{or} \qquad J(P,C) = e(t)^2 \qquad (7)$$

where e(t) is an error function depending linearly upon the system output y(t), matrix H and vector $\operatorname{grad}_c J$ become [9], [11]

$$H = [\overline{V(t)\;V(t)'} + \overline{e(t)\;W(t)}]^{-1}_{.} \qquad (8)$$

$$\operatorname{grad}_c [J(P,C)] = \overline{e(t)\;V(t)} \; ,$$

where $V(t) = \{v_n(t)\}$ is sensitivity vector of first order sensitivity functions,

$$v_n(t) = \partial y(t)/\partial C_n \; ,$$

$W(t) = \{w_{nm}(t)\}$ is (NxN) matrix, whose elements are second order sensitivity functions [12],

$$w_{nm}(t) = \frac{\partial^2 y(t)}{\partial c_n \partial c_m}$$

and (6) denotes vector transposition.

In model reference adaptive systems, it often occurs that the reference model is such that

$$e(t,P^o,C^o) \to 0 \qquad (9)$$

where C^o are optimum values of adjustable controller parameters, corresponding to the nominal values P^o of variable plant parameters, i.e.

$$J_{min} = J(P^o,C^o) = \min_C [J(P^o,C)]. \tag{10}$$

Consequently, neglecting the second term on the right side of (8), the matrix H may be approximated by

$$H = [\,\overline{V(t)\ V(t)'}\,]^{-1} \quad . \tag{11}$$

As relations (8) and (11) use sensitivity functions, it is evident that they are valid only for slow parameter variations. Furthermore, as the form of sensitivity functions $v_n(t)$ depends upon the P and C it may be concluded that using the expressions (8) and (11) for large parameter variations would require continuous calculation and readjustment of matrix H, which is hardly acceptable for implementation in adaptive systems.

Adaptive System with Fixed Matrix H

In extending these ideas to large parameter variations and to enable realization of practically acceptable adaptive systems, it is necessary to solve the following two problems. The first problem is that the controller with adjustable parameters C should be selected in such a manner to satisfy the condition of complete compatibility in the large, discussed in [1], [13]. This condition may be defined by: The controller (1) is completely compatible in the large to the given plant if, for every $P \in D_p$, in the domain D_p of expected plant parameter variations, there exists a vector relation

$$C = F\ (P) \quad , \tag{12}$$

implementable by the adaptive system, such that

$$J[P,F(P)] = J[P^o,C^o] \ . \tag{13}$$

Second problem consists in dependence of $v_n(t)$ on system parameters P and C. If we assume that adaptive system, minimizing the performance criteria (7) implements the relation (12), it is sufficient to analyze the dependence

$$v_n[t,P,F(P)]$$

along the adaptation trajectories (12) only.

From structural analysis of the linear, multiloop system of the type treated here (Fig. 2), where plant and controller parameters are gain factors of the isolated blocks, it follows that controller sensitivity functions $v_n(t)$ along the adaptation trajectories (12) behave as

$$v_n(t) = \frac{1}{c_n}\ v_n^{*}(t) \quad , \tag{14}$$

where $v_n^{*}(t)$, known as logarithmic sensitivity function, is not dependent upon the point P,F(P), i.e.

$$v_n^{*}[t,P^o,C^o] = v_n^{*}[t,P,F(P)]. \tag{15}$$

According to (14) vector $V^{*}(t)$ may be expressed as

$$V^*(t) = \text{diag}\ C\ V(t) = G\ V(t) \qquad (16)$$

where $G = \text{diag}\ C = \{\delta_{nm} c_n\}$, $(n,m = 1,2,\ldots N)$ is NxN diagonal matrix with c_n in main diagonal, and δ_{nm} is the kronecker symbol.

Determination of the Transformed Algorithm

The property (16) may be successfully applied in attempting to cover whole domain D_p by a fixed matrix H^*, rather than by a variable H depending upon P and F(P).

Using (8), (11), and (16), the matrix H and the vector grad_c J become

$$H = G\ [\ \overline{V^*(t)\ V^*(t)'}\ ]^{-1} \cdot G = G \cdot H^* \cdot G\ , \qquad (17)$$

$$\text{grad}_c\ J = G^{-1}\ \overline{e(t)\ V^*(t)}\ .$$

Introducing

$$C^* = \ln C,$$

and using (17), the transformed algorithm becomes finally

$$\dot{C}^* = -\ [\ \overline{V^*(t)\ V^*(t)'}\ ]^{-1} \cdot \overline{e(t)\ V^*(t)} \qquad (18)$$

where

$$\dot{C} = G\ \dot{C}^*\ .$$

As the logarithmic sensitivity functions $v^*(t)$ in our case do not depend upon the trajectories (12), it may be concluded that the matrix

$$H^* = [\ \overline{V^*(t)\ V^*(t)'}\ ]^{-1} \qquad (19)$$

may be calculated in advance (off-line) and included into the adjusting mechanism as fixed time-constant matrix. Block diagram of the adaptive system, based upon the algorithm (18) is given on Fig. 1. Dotted lines represent off-line adjustments.

The blocks denoted by S, R and SM are the system, the reference model and the sensitivity models, respectively. The reference model generates reference output $y^{o}(t)$, while the sensitivity model generates the sensitivity functions $v_n^*(t)$. According to the results obtained in [1], it was possible to fix parameters P and C in the sensitivity model at their nominal or optimal values. The fact that all logarithmic sensitivity functions $v_n^*(t)$ in the case of linear systems are available from one sensitivity model which is identical to the model of the system, [14], [15], further simplifies practical implementation of the multiparameter adaptive system.

Strictly speaking the obtained algorithm is applicable only for the slow variations of plant parameters. However, computer results have shown that deteriorations in the sensitivity functions, due to high speed variations of P and C, do not affect appreciably the parameter trajectories and adaptation process. Consequently, this permits application of the transformed algorithm (18) in the synthesis of high-speed multiparameter adaptive systems.

Example

In order to illustrate the improvements obtained using the algorithm (18) we have taken a three-parameter adaptive system controlling a linear fourth order plant. A block diagram of the system is represented on Fig. 2. It is shown that the selected controller

(c_1, c_2 and c_3 are adjustable)

satisfies the condition of complete compatibility in the large [13], given by (13). Variable plant parameters are p_1, p_2 and p_3. Nominal and optimal values of plant and controller parameters are

$$p_1 = p_2 = p_3 = c_1 = c_2 = c_3 = 1 \quad . \qquad (20)$$

The controller parameters C have been adjusted by the algorithm (18), i.e. by

$$\dot{C} = - \operatorname{diag} C \, H^{*} \cdot e(t) \cdot V^{*}(t) \qquad (21)$$

where H^{*} is a constant coefficient matrix, which is in our case given by

$$H^{*} = \begin{vmatrix} 0.5782 & -0.0272 & -0.0973 \\ -0.0272 & 0.1050 & 0.0156 \\ -0.0973 & 0.0156 & 0.0561 \end{vmatrix} \quad . \qquad (22)$$

Fig. 3 gives sample of results obtained by analog computer simulation of this three parameter adaptive system. As plant perturbations, we have varied parameter p_3 according to Fig. 3a in the range of 800%. Of course, the selected values of plant parameter variation are such that without the adaptation the system responses would become highly unstable. Fig. 3b corresponds to application of the transformed algorithm (18). To appreciate the advantages of algorithm (18) over the classical algorithm (2) it should be mentioned that application of (2), i.e.

$$\dot{c}_n = - h_n \cdot \partial J(P,C)/\partial c_n \quad (n = 1,2,3) \qquad (23)$$

where

$$h_n = h$$

has led in our case to either unstable or very slow adaptation. However, the only possibility of obtaining stable operation was to decrease considerably the value of scalar h only in the first equation, i.e.

$$h_3 = h_2 = 10 h_1 \quad . \qquad (24)$$

However this situation corresponds to adaptive system which is more similar to two-, rather, to three-parameter adaptive systems. System responses in this case are given on Fig. 3c. Comparing Fig. 3b and Fig. 3c it may be concluded that as plant parameters variations become larger, the improvements obtained by algorithm (18) are more remarkable. Fig. 4b shows how the adaptive system, based upon the transformed algorithm (18), adjusts controller parameters c_1, c_2, c_3, when plant parameter p_3 varies according to Fig. 4a.

The similar results have also been obtained in the cases of variations of parameters p_1 and p_2.

Conclusion

In this paper it is shown that the well known iteration algorithms, widely used in parameter optimization and system identification, may be successfully applied in synthesis of multiparameter adaptive system. It is also shown that the large and high speed variations of plant parameters may be successfully compensated if the condition of compatibility of the adaptive controller is fulfilled.

Acknowledgment

The author wishes to express his appreciation to Prof. P. V. Kokotović for his constructive criticism and advice.

References

[1] P. V. Kokotovic, J. Medanic, M. Vuskovic and S. P. Bingulac, Sensitivity Method in the Experimental Design of Adaptive Control Systems, Proc. III-rd IFAC Congress, London, 1966.

[2] P. Eykhoff, Some Fundamental Aspects of Process Parameter Estimation, IEEE Trans. on Automatic Control, Vol. AC8, Oct. 1963, pp. 347-357.

[3] M. Margolis and C. Leondes, On the Theory of Adaptive Control Systems, the Learning Model Approach, I st IFAC Congress, Moscow, 1960.

[4] D. Donaldson and C. Leondes, A Model References Paramerer Tracking Technique for Adaptive Control Systems, I - The Principle of Adaptation, IEEE Trans. on Application and Industry, Vol. 82, Sept. 1963, pp. 241-252.

[5] R. E. Andeen, A General Adaptive Approach Based upon Power Density Spctrum, Proc. III-rd IFAC Congress, London, 1966.

[6] R. E. Butler and E. V. Bohn, An Automatic Identification Technique for a Class of Nonlinear Systems, IEEE Trans. on Automatic Control, Vol. AC11, April 1966, pp. 292-296.

[7] C. Csapodi, Discussion on the Paper cited in [1], Proc. III-rd IFAC Congress, London, 1966.

[8] N. N. Puri and C. N. Weygandt, Multivariable Adaptive Control Systems, Proc. Joint Automatic Control Conf. New York, 1962.

[9] S. P. Bingulac, On the Improvement of Gradient Method, IBK publication, No. 450, May 1966. Beograd.

[10] P. V. Kokotovic, Note on an Improvement of the Simultaneous Gradient Method for Parameter Optimization, Pupin Control Lab. Report, June 1966.

[11] L. E. McBride and K. Steiglitz, Iterative Methods for Systems Identification, Princeton University, Communication Lab. Technical Report No. 15, June 1966, Princeton, N.J.

[12] S. P. Bingulac, Simultaneous Generation of the Second Order Sensitivity Functions, IEEE Trans. on Automatic Control, Vol. AC11, July 1966.

[13] S. P. Bingulac, On the Compatibility of Adaptive Controller, Fourth Allerton Conf. University of Illinois, Urbana, Illinois, 1966.

[14] P. V. Kokotovic, Simultaneous Computation of Sensitivity Functions, Pupin Control Lab. Report, July 1962, also published in R. Tomovic, Sensitivity Analysis of Dynamic Systems, pp. 57-78, McGraw Hill Book, New York, 1965.

[15] P. V. Kokotovic, Sensitivity Points Method in Analysis and Optimization of Linear Control Systems, Automatic and Remote Control, No. 12, Dec. 1964, Moscow.

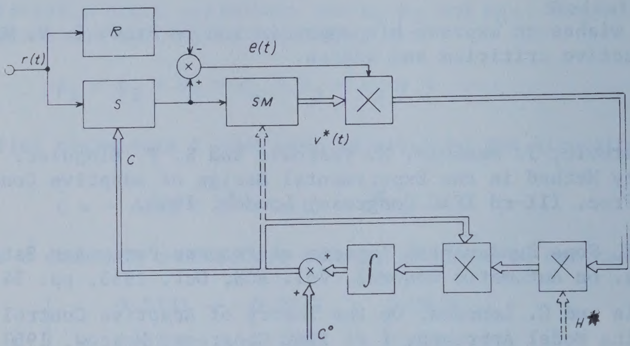

R
r(t)
e(t)
S
SM
v*(t)
C
∫
C°
H*
+
-

Fig.1

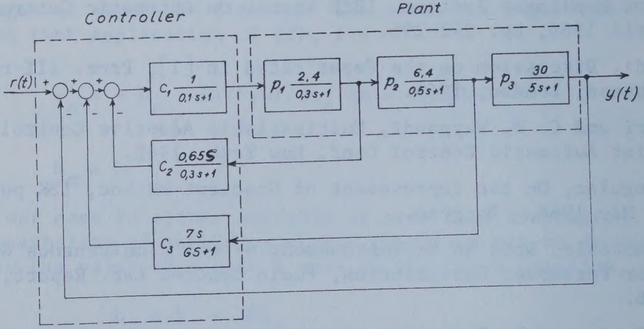

Controller
Plant
r(t)
$C_1 \frac{1}{0,1s+1}$
$P_1 \frac{2,4}{0,3s+1}$
$P_2 \frac{6,4}{0,5s+1}$
$P_3 \frac{30}{5s+1}$
y(t)
$C_2 \frac{0,65S}{0,3s+1}$
$C_3 \frac{7s}{6s+1}$
+
-

Fig 2

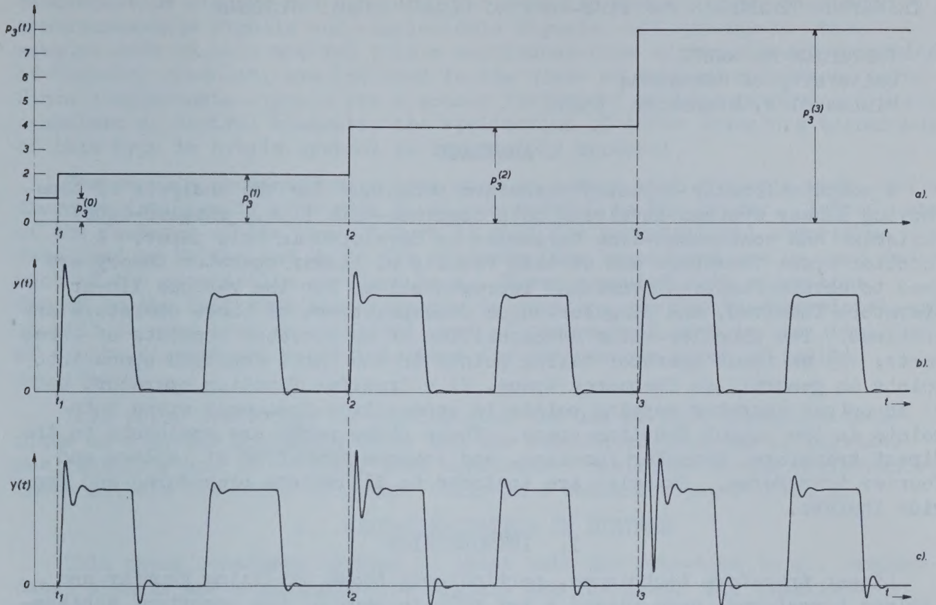

$p_3(t)$
7
6
5
4
3
2
1
0
$p_3^{(0)}$
$p_3^{(1)}$
$p_3^{(2)}$
$p_3^{(3)}$
a).
t_1
t_2
t_3
t
$y(t)$
0
b).
$y(t)$
0
c).

Fig. 3

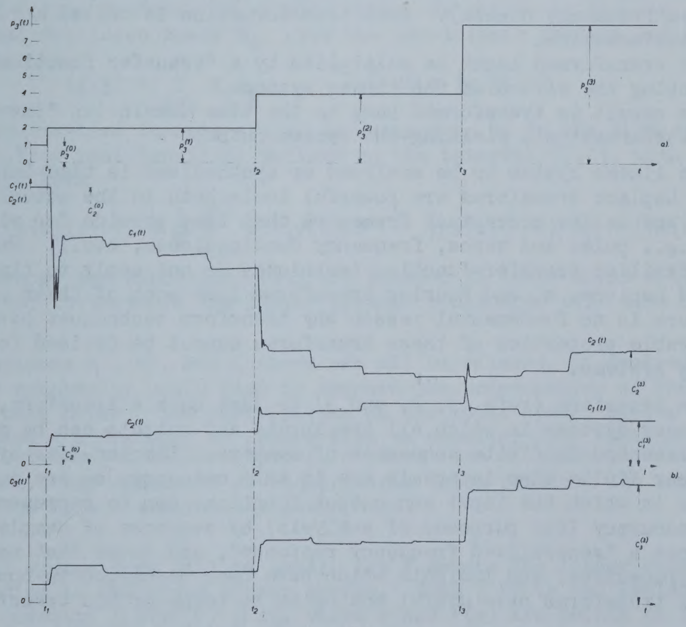

$p_3(t)$
7
6
5
4
3
2
1
0
$p_3^{(0)}$
$p_3^{(1)}$
$p_3^{(2)}$
$p_3^{(3)}$
a)
t
t_1
t_2
t_3
$c_1(t)$
$c_2(t)$
$c_2^{(0)}$
$c_1(t)$
$c_2(t)$
$c_2^{(3)}$
$c_1(t)$
$c_1^{(3)}$
$c_2(t)$
$c_2^{(0)}$
0
b)
$c_3(t)$
$c_3^{(3)}$
0
t

Fig 4

A TRANSFORM TECHNIQUE FOR TIME-VARYING LINEAR HYBRID SYSTEMS

FREDERICK M. WALTZ
University of Minnesota
Minneapolis, Minnesota 55455

ABSTRACT

A computationally-oriented transform technique for the analysis of time-varying linear systems involving both discrete-time (i.e., sampled-data) variables and continuous-time variables is developed in this paper. A function-space framework and certain results of linear operator theory are used to obtain finite-dimensional representations for the various linear operators involved, and singular-value decompositions of these operators are obtained. The singular-value decomposition of an operator consists of three parts: 1) an input operator taking points in the input function space into points in generalized frequency space, 2) a transfer function operator, and 3) an output operator mapping points in generalized frequency space into points in the output function space. These three parts are analogous to the direct transform, transfer function, and inverse transform of Laplace and Fourier transforms. Examples are included to illustrate procedures and provide insight.

1. INTRODUCTION

Linear transform techniques, particularly those utilizing Fourier and Laplace transforms, have played a key role in many of the important achievements in electrical engineering as well as in other engineering and scientific fields. The underlying pattern, shared by all the linear transform techniques being considered here, is as follows:

a) The input to the linear system in question is transformed from the time domain to the transform domain (e.g., frequency domain, complex frequency domain). This transformation is called a "direct" transformation.
b) The transformed input is multiplied by a "transfer function" representing the effect of the linear system.
c) The result is transformed back to the time domain (an "inverse" transformation), yielding the system output.

When the linear system to be analyzed or synthesized is time-invariant, Fourier and Laplace transforms are powerful tools both in the actual solution of problems and in the conceptual framework that they provide for viewing problems (e.g., poles and zeros, frequency domains ideas, etc.). Unfortunately, the familiar transfer-function techniques do not apply to time-varying systems, and Laplace, z, and Fourier transforms lose much of their usefulness. However, there is no fundamental reason why transform techniques having many of the desirable properties of these transforms cannot be devised for linear time-varying systems.

Naylor's transform (refs. 1, 2, and 3) is just such a transform, applicable to linear systems in which all the inputs and outputs can be adequately represented by finite sequences of numbers. Sampled-data systems operating over finite time intervals are in this category, as are continuous time systems in which the input and output functions can be represented with sufficient accuracy (for purposes of analysis) by sequence of samples. Naylor defines a "generalized frequency response", and shows that many of the ideas, procedures, and insights which have been developed in connection with Fourier transforms have useful analogies in terms of his transform.

The present paper develops similar (and compatible) transforms for certain time-varying linear systems which involve both continuous-time and discrete-time signals and systems. For example, systems involving both continuous-time signals and sampled-data signals, and systems in which sampled-data signals are fed into a continuous-time system through some kind of "holding circuit", are included in the class of systems considered here. Since sampled-data signals are a common feature of systems involving digital computers as control elements, the application of these transform techniques of this type to hybrid systems is immediately apparent.

The results presented here have much in common with Naylor's work, and a thorough familiarity with ref. 1 will be helpful to the reader. Indeed, one of the features of the present work is that the computational algorithms developed by Naylor and Waltz for discrete-time transforms can be used for much of the computation involved here. These computational algorithms depend on the fact that the operators in question are finite-dimensional, and it will be seen that an essential part of the present work is the obtaining of finite-dimensional representations for the various types of systems considered.

This paper deals only in passing with the use of this transform technique as a system synthesis tool, and the associated problems of physical realizability and of deciding what constitutes desirable system performance. These matters are dealt with more fully in a companion paper (ref. 4).

2. CHARACTERIZATION OF SYSTEMS

This paper considers systems in which both discrete-time (e.g., sampled-data) signals and continuous-time signals are involved. For simplicity it will be assumed that a) all systems and blocks within systems have one input variable and one output variable; b) all systems operate over the same finite interval [0,T]; and c) unless otherwise stated, all sampled-data signal sequences consist of N instantaneous samples of some continuous-time waveform taken at the equally-spaced times, $0, \tau, \ldots, (N-1)\tau$, where $\tau = T/N$. The N-component sampled-data sequences will be regarded as N-component column vectors, or, equivalently, as points in the real N-dimensional Euclidean space E_N, with the usual inner product and norm

$$(\underline{x},\underline{y}) = \sum_{i=1}^{N} x_i y_i \qquad \|\underline{x}\| = (\underline{x},\underline{x})^{1/2} \tag{1}$$

Continuous-time functions will be assumed to be in the Hilbert space $L_2 = L_2(0,T)$ of real functions defined on the interval [0,T], with inner product and norm

$$\langle x(t),y(t)\rangle = \frac{1}{\tau}\int_0^T x(t)y(t)dt \qquad \|x(t)\| = \|x\| = [\langle x(t),x(t)\rangle]^{1/2} \tag{2}$$

Only sequences and functions which are bounded in these norms will be considered, and furthermore the linear systems considered will also be bounded in norm.*

Assumptions a), b), and c) above can all be relaxed, if desired, but the resulting complexity would tend to obscure the presentation of the basic ideas of this paper. The restrictions on the input space, system norm, and output space are designed to insure that we consider only continuous linear mappings between two Hilbert spaces, at least one of which is finite dimensional. Any other set of requirements which produces the same results will suffice from a theoretical standpoint.

---

*The usual system norm will be used: Let P denote the linear operator corresponding to the given linear system. Then $\|P\|$ is the maximum over all x of the quantity $\|P(x)\| / \|x\|$, where x and P(x) are points in E_N or L_2, depending on the type of system being considered, and where the appropriate norms are used in each case.

The four types of systems to be considered here will be referred to as systems of types D-C, C-D, D-D, and C-C, where in each case the first letter refers to the form of the input variable (D for discrete-time, C for continuous-time), and the second letter gives the corresponding information about the output variable. Systems made up of subsystems of these four types interconnected in various ways (loops, strings, etc.) will be investigated here.

Clearly, systems of these types cannot be interconnected arbitrarily. For example, a system of Type C-D must receive its input either from a system of Type C-C or D-C (or some combination of such systems), and its output can be fed only into systems of Types D-D or D-C.

The following notational conventions are used in this paper: All vectors will be understood to be column vectors, with row vectors being denoted as transposed column vectors. A prime denotes transposition. Vectors will be denoted by underlined symbols [e.g., $\underline{u}$ and $\underline{v}(t)$] and the components of the vectors by the same symbols with subscripts 1 through N (e.g., $u_1, u_2, \ldots, u_N$). A vector function $\underline{v}(t)$ will be understood to be an M-tuple or N-tuple of functions in L_2, and may also be regarded as a time-varying vector with values in E_M or E_N at each moment in the interval $[0,T]$. Quantities involving the inner product symbols $\langle , \rangle$ and vector functions are obtained by carrying out the inner product component by component. Thus, for example, $\langle \underline{f}(\xi), x(\xi) \rangle$ is a column vector with components $\langle f_i(\xi), x(\xi) \rangle$ and $\langle \underline{f}(\xi), \underline{g}'(\xi) \rangle$ is an NxN matrix with components $\langle f_i(\xi), g_j(\xi) \rangle$.

Operator Representations

The first lines of the boxes in the second column of Table 1 give representations for the four types of operators which are valid over the whole input and output spaces (E_N and/or L_2). These representations can be obtained either experimentally or by theoretical analyses of mathematical models of the systems. For Type D-C systems, the characterizing vector function $\underline{p}(t)$ has as its i-th component $p_i(t)$, the system output which results from the application of the i-th "natural" input basis sequence; that is, the input sequence for which the i-th component is unity and all other components are zero. For Type C-D systems, $p_i(t)$ is τ times the output of the system at the i-th sampling time, $(i-1)\tau$, which results from a unit impulse applied at time t. For Type D-D systems, the i-th column of the matrix P is the output sequence which results from the application of the i-th "natural" input basis sequence. Finally, the function $p(t,\xi)$ used in the representation of Type C-C operators is just τ times the unit impulse response of the system.

When systems of these four types are interconnected, it will often be the case that a given system cannot have certain functions or sequences as input, since they lie outside the (finite-dimensional) range of the system which feeds the system in question. Similarly, we may not be concerned with outputs from a given system which lie in the null space of the system which it feeds. We may therefore be interested in some cases in representing operators only over some subspaces of the input space or output space, or perhaps both. Such restricted representations are easily obtained from the unrestricted representations, as shown in the remaining columns of Table 1. In each case the restricted input subspaces are specified as being spanned by the (orthonormal) columns of the matrix A or the (orthonormal) components of $\underline{a}(t)$. The restricted output subspaces are similarly spanned by the columns of B or components of $\underline{b}(t)$. Note from this table that, while the unrestricted representation for Type C-C operators is infinite-dimensional, in general, all the restricted representations are finite-dimensional, in that each involves only a finite number of vectors or functions. Since this paper considers Type C-C systems only as subsystems in larger systems which also involve subsystems of the other three types, in all the cases discussed here it will be possible to use finite-dimensional representations for Type C-C systems, when they occur. This point is illustrated by the example shown in Fig. 1a), in which a Type C-C system has as its input the output of a Type D-C system. Figure 1b) shows the same system in terms of function spaces and linear operators, where here and in the diagrams to follow a rectangular figure is used to represent a finite-dimensional space (or subspace) and an amorphous figure is used to represent an infinite-dimensional space (or subspace). The quantity $F(E_N)$ is defined as the set of all possible outputs from F using inputs from E_N; other quantities involving operators operating on spaces or subspaces are defined analogously.

The important point to note here is that even though the output z(t) of G lies in the infinite dimensional space L_2, the actual outputs that can be achieved from G, using inputs from F, lie in the N-dimensional (or smaller) subspace $G(F(E_N))$ of L_2. Thus G can be regarded as a transformation from one finite-dimensional space to another and many of the difficulties encountered in working with infinite-dimensional spaces can be avoided.

3. DECOMPOSITIONS AND INVERSES OF OPERATORS

Singular-Value Decompositions

The concept of the singular values of an operator is an extension of the idea of the norm of an operator. In fact, the largest singular value (call it λ_1) is just the norm of the operator over the domain in question. Associated with this value λ_1 is some input x_1 in the domain for which the maximum involved in the norm is actually taken on. Any such input will suffice. The next largest singular value λ_2 (which may equal but cannot exceed λ_1) is then the norm of the operator over that part of the domain orthogonal to x_1, and associated with λ_2 will be some other input x_2 (orthogonal to x_1) for which the maximum is taken on. The third singular value λ_3 is then the norm of the operator over that part of the domain orthogonal to x_1 and x_2. And so on.

If we denote the dimension of the range of the operator by M, then there will be precisely M nonzero singular values and M associated mutually orthogonal inputs. (For the cases considered here, the range is always of dimension N or less, so $M \leq N$ here.) The M associated inputs $x_1, x_2, \ldots, x_M$ form an orthogonal basis for the significant space* of the operator, and the (normalized) operator outputs $y_1, y_2, \ldots, y_M$; $y_i = 1/\lambda_i \, P(x_i)$, corresponding to these inputs form an orthogonal basis for the range of the operator. The singular value decomposition of an operator is just the arrangement of these singular values, appropriately normalized input basis functions (or vectors), and appropriately normalized output basis function (or vectors) into convenient vector-matrix form, making use of L_2-space inner product notation where needed, as shown on the second line of each box in Table 1. The expressions in brackets in each box give the vector function or matrix which must be decomposed, followed by a statement of which of the three given procedures may be used to carry out the decomposition. Note in passing that there is in general no simple relationship between the unrestricted decomposition of an operator and its restricted decompositions.

The sequential procedure outlined above for the determination of singular values could actually be used as the basis for a computational procedure for obtaining singular value decompositions, and such an approach might be very convenient if M were large but only some relatively small subset of the M nonzero singular values were large enough, compared to λ_1, to merit computation from an engineering standpoint. However, the singular value decomposition procedures used in previous works involving matrix decompositions (refs. 1, 2, and 3) make use of the fact that the singular values of P are the positive square roots of the eigenvalues of the matrix P'P, and an input basis for P is any set of orthonormal eigenvectors for P'P. Since standard computer subroutines are available for determining the eigenvalues and eigenvectors of a matrix, this was the approach taken. The present paper makes use of this same approach, so that the computer programs already available can be used. Decomposition procedures involving the direct mechanization of the above definition will be discussed in a later paper.

---

*The significant space of an operator is, by definition, the subspace of the input space formed by taking the orthogonal complement of the null space of the operator and adding to it the zero element.

By whatever method they are obtained, these singular value decompositions form the basis for the transform technique presented here. In each case the decomposition of an operator P can be viewed as a resolving of P into three steps:

a) Operation on the input $\underline{x}$ or x(t) by means of the operator X' or $\langle \underline{x}(t),*\rangle$, which constitutes a transformation of the input from the input space (i.e., time domain) into what Naylor (refs. 1,2,3) calls the "generalized frequency domain". The columns of X or the components of $\underline{x}(t)$ form an orthonormal set spanning the significant space (or the restricted input subspace) of P, and the result $X'\underline{x}$ or $\langle \underline{x}(t),x(t)\rangle$ is a column vector, the components of which give the expansion of the input in terms of the spanning set (analogous to the coefficients of a Fourier series expansion, except in this case there are only a finite number of coefficients).

b) Multiplication of each component of the expansion (i.e., transformed input) by the system gain λ_i for that component. The diagonal matrix Λ can thus be regarded as the transfer function (or, alternatively, as the generalized frequency response) of the system P.

c) The formation of a linear combination of the output spanning set, with the coefficient being those determined from steps a) and b). This is analogous to taking the inverse transform, and will be so regarded.

Inverse Operators

The operators shown in Table 1 may not have inverses; that is, for a given operator P there may be no operator P^{-1} such that $P^{-1}(P(x)) = x$ for every x in the domain of P. In fact, operators of Type C-D can never have inverses, since the dimension of the range of the inverse would have to be infinite, but the domain is finite dimensional. For operators of the other types, it is often the case in practice that inverses do not exist. However, if we restrict P to be an operator mapping points in its significant space (or some restricted subspace thereof) into points in its range (or the correspondingly restricted output subspace), then this operator will have an inverse (actually a pseudo-inverse) mapping points in the range (or restricted output subspace) of P into points in the significant space (or restricted input subspace) of P. It is in this sense that the term "inverse" and the superscripted (-1) are used here.

The inverse P^{-1} of a given operator P can be obtained easily, once the singular value decomposition of P is known. The input basis of P becomes the output basis of P^{-1}, and vice versa, and the transfer function of P^{-1} is the inverse of that of P. For example, if $P = \underline{b}'(t)QA' = [\underline{b}'(t)Y]\Lambda[X'A']$ then $P^{-1}(*) = [XA]\Lambda^{-1}\langle Y'\underline{b}(\xi),*\rangle$. The inverse transfer function Λ^{-1} always exists (because zero singular values have been deleted from Λ) and is easily obtained from Λ simply by computing M reciprocals. This type of inversion procedure extends even to operator representations which are not in the form of singular value decompositions. For example, let $P(*) = \underline{b}'(t)Q\langle \underline{a}(\xi),*\rangle$, where Q is any non-singular matrix and $\underline{a}(t)$ and $\underline{b}(t)$ are orthonormal. Then $P^{-1}(*) = \underline{a}'(t)Q^{-1}\langle \underline{b}(\xi),*\rangle$, as is easily verified by substitution and by use of the facts that the quantity $\langle \underline{b}(\xi), \underline{b}'(\xi) \rangle$ is the identity matrix and the quantity $\underline{a}'(t)\langle \underline{a}(\xi), * \rangle$ maps points in the subspace of L_2 spanned by the components of $\underline{a}(t)$ into themselves.

4. ANALYSIS OF SYSTEMS USING THE TRANSFORM TECHNIQUE

One use of the transform technique presented above will now be illustrated by analyzing two examples with it. The first example shows some of the ways in which the operators and transforms can be manipulated to achieve desired ends, and the second carries the process through to a numerical result in order to illustrate the kind of results that can be expected.

Example 1. A Feedback Control System of Type D-D

Assume that the Type D-D feedback control system shown in Fig. 2 is given. Let the unrestricted decomposition of G be $G = \underline{g}'(t) = \underline{y}_G'(t) \Lambda_G X_G'$ and let the decomposition of F with input subspace restricted to the subspace spanned by the components of $\underline{y}_G(t)$ be $F(*) = Y_F \Lambda_F \langle X_F' \, \underline{y}_G(\xi), * \rangle$. The operator P characterizing the overall system can then be derived using these decompositions as a basis:

$$\underline{c} = F(m(t)) = F(G\underline{e}) = F(G(\underline{r}-\underline{c})) = F(G\underline{r}) - F(G\underline{c})$$

$$\underline{c} + F(G\underline{c}) = F(G\underline{r})$$

$$[I + F(G)]\underline{c} = F(G)\underline{r}$$

$$\underline{c} = [I + F(G)]^{-1} F(G)\underline{r}$$

Hence $P = [I + F(G)]^{-1}F(G)$, $P = [I + Y_F \Lambda_F X_F' \Lambda_G X_G']^{-1} \, Y_F \Lambda_F X_F' \Lambda_G X_G'$ where the fact that $\langle \underline{y}_G'(\xi), y_G'(\xi) \rangle$ is the identity matrix has been used to simplify the final expression.* Thus, the analysis of the overall behavior of this system can be carried out in terms of matrix manipulations, even though continuous-time functions are involved at one point in the system. The singular value decompositions of P can be obtained using Procedure 1 of Table 1 on the matrix P.

Example 2. A Simple Numerical Example

As a second example, consider the simple Type D-C feedback control system shown in Fig. 3, in which the forward loop (characterized by the operator F) consists of an ideal holding circuit (that is, one whose output is held equal to the previous sample value until the next sample arrives, at which time it switches instantaneously to the new value), feeding an ideal amplifier with time-varying gain t/T. The feedback loop consists of an ideal integrator of unit gain, the output of which is sampled instantaneously at time $0, \tau, \ldots, (N-1)\tau$ to produce the output sequence $\underline{b}$.

For this system, we have $c(t) = F\underline{r} - FG(c(t)) = [I(*) + FG(*)]^{-1}(F\underline{r})$ so that the overall system operator P is given by $P = [I(*) + FG(*)]^{-1}(F)$ which, upon substitution of the unrestricted decomposition $\underline{y}_F'(t) \Lambda_F X_F'$ for F and the restricted representation $H\langle \underline{y}_F(\xi), * \rangle$ for $G(*)$ (input restricted to the range of F), becomes

$$P = [\underline{y}_F'(t)\langle \underline{y}_F(\xi), * \rangle + \underline{y}_F'(t) \Lambda_F X_F' \, H\langle \underline{y}_F(\xi), * \rangle]^{-1} \, (\underline{y}_F'(t) \Lambda_F X_F')$$

$$P = \underline{y}_F'(t) \, [I + \Lambda_F X_F' H]^{-1} \, \Lambda_F X_F'$$

---

*It is stated without proof that, if systems which are not physically realizable and Type C-D systems which are pure instantaneous unit-gain samplers with no integration or time delay involved are excluded, then the inverse matrix in this expression and in the similar expression to follow will always exist.

It is easily shown that for this system $X_F = I$, that the diagonal elements λ_{Fi} of Λ_F are $\lambda_{Fi} = [(3i^2 - 3i + 1)/3N^2]^{1/2}$, that the component $y_{Fi}(t)$ of $\underline{y}_F(t)$ are $y_{Fi}(t) = t/\lambda_i$ on the interval $((i-1)\tau, i\tau]$ and zero elsewhere, and that the components h_{ij} of H are $h_{ij} = (2j-1)\tau^2/2\lambda_j N$ for $j \leq i-1$ and zero of all other values of i and j. These results, along with the corresponding decomposition for the closed loop system operator P, are shown in Fig. 4 for the simple case N = 4. (A small value of N has been used here deliberately in order that the reader may repeat the numerical calculations himself without undue difficulties if he so desires. Even for such a small value of N the results are not useless, however, since they allow one to study the "start-up transients" of such a system.)

Further Discussion

In the above examples it has been shown that the analysis of the system in question can be reduced to the manipulation of singular-value decompositions of linear operators characterizing the various subsystems of the overall system, and that the actual computations required for numerical solution of these problems can be carried out in a routine way using existing computational techniques and subroutines. For more complicated examples the work involved will be greater, but the same procedures can be followed.

The interpretation of the results of an analysis of this type can be approached in various ways. The most direct way is in terms of the definition of the singular value decomposition: In Example 2, for instance, every output that can be obtained on the interval [0,4] must be some linear combination of the four output basis functions shown in Fig. 4. Furthermore, the last output basis function $y_4(t)$ is the one for which the system has the highest gain, namely λ_4; if one applies $\underline{x}_4$, the last input sequence shown, then the resulting output will be λ_4 times $y_4(t)$. Any other input sequence will yield an output function of smaller norm.

Further analysis along this line can be carried out. However a primary aim of this paper has been to establish these results as a generalization of the previous work of Naylor and Waltz (refs. 1, 2, and 3), so that the many ideas and discussions in these references can be carried over to the area of "hybrid" systems. Therefore, further discussion of the interpretation of results, and of the analogies between these results and other and more familiar transform techniques, are omitted here as being repetitious of discussions already presented in the stated references.

5. ADVANTAGES AND LIMITATIONS

The value of the transform technique presented here lies in the fact that it, along with the linear operator notation used here, provide a straightforward method of formulating and solving problems involving linear time-varying hybrid systems. In addition, such an approach may prove to be valuable as a frame of reference for considering such problems, even if actual numerical results are not needed or wanted, analogous to the way that concepts such as impedance and frequency response are valuable in linear time-invariant problems, even when actual numbers are not used.

There are two limitations involved in this approach at present, however: 1) the potential user of the technique may not possess the insight to interpret his results properly; and 2) for large problems (i.e., for values of N larger than 50 or so) the computation time on the digital computer becomes excessive when present subroutines are used. The first limitation is shared by most if not all techniques when they are first introduced, and should be remedied with time. Some progress has already been made (see ref. 3) in this direction with Naylor's transform (i.e., the transform presented here for Type D-D systems), and many of the same ideas carry over to the other three system types considered here.

The second limitation noted above is a problem shared with Naylor's approach, the major amount of time in each case being that used by the eigenvalue-eigenvector subroutine used to compute the singular values. Various solutions to this problem are possible: development of singular-value subroutines adapted to the special structure of low triangular matrices; approximation techniques; a procedure for sequentially determining singular-values (in order of decreasing magnitude) and stopping the procedure when the values get smaller than some preset level (i.e., when essentially all of the system gain has been accounted for); etc. These and other possibilities are being or will be investigated.

It should be pointed out here, however, that in some cases the transforms used here will not suffer from this second limitation as much as Naylor's transform would when applied to the same problem, or, stated positively, a value of N smaller than 50 or so may often suffice for the problems considered here. This is because, in Naylor's approach, the value of N must be chosen large enough to give an adequate approximation to the time waveforms involved, and may in some cases be considerably greater than the number of sampling instants. In the approach given here, N is the number of sampling instants, and the accurate representation of time waveforms is separated entirely from the choice of N. Waveform representation may be accomplished in various ways: on an analog device (e.g., tape recorder), or on a digital computer by using a finer time subdivision than the sampling interval. Such "fine-grain" representation increases the integration times in inner-product computations, but it does not increase the size of the matrix (NxN) which must be decomposed, and it is the matrix decomposition which takes most of the computing time as N get large.

Another area that is currently being investigated (see ref. 4) is the use of this technique for synthesizing hybrid systems; that is, given some parts of the system (e.g., the "plant") and some criteria of system behavior, choose the remaining parts of the system (e.g., the "controller") so as to satisfy the design criteria. This area also involves a study of the physical realizability of such systems; i.e., what forms do the system operators take when the systems are required to be physically realizable.

The net result of the present work and the various extensions of it suggested above will be, it is hoped, a routine method for formulating and solving a wide variety of problems involving continuous-time and discrete-time variables.

REFERENCES

1. A. W. Naylor, "Generalized Frequency Response Concepts for Time-Varying, Discrete-Time, Linear Systems," IEEE Transactions on Circuit Theory CT-10, No. 3, September 1963.

2. A. W. Naylor, "A Transform Technique for Multivariables, Time-Varying, Discrete-Time, Linear Systems," Automatica 2, 211-234, 1965.

3. A. W. Naylor and F. M. Waltz, "On the Synthesis of Time-Varying, Discrete-Time, Linear Feedback Control Systems with a New Transform Technique," to be published in IEEE Transactions on Automatic Control, October 1966.

4. F. M. Waltz, "On the Synthesis of Time-Varying Linear Hybrid Systems with a New Transform Technique," to be published.

Table I Representations and Decompositions of Unrestricted and Restricted Operators

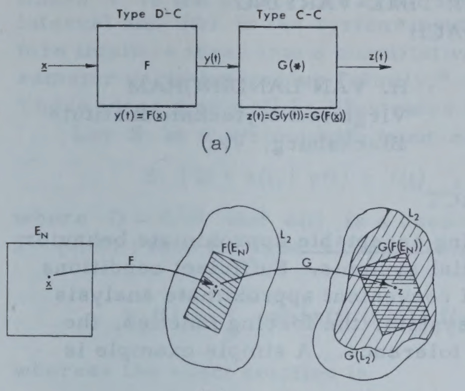
System Type	Unrestricted Operator	Restricted Input Subspace	Restricted Output Subspace	Restricted Input and Output Subspaces
D-C	$y(t) = \underline{p}'(t)\underline{x}$ $= \underline{y}'(t)\Lambda X'\underline{x}$ [$\underline{p}(t)$: Procedure 2]	$y(t) = \underline{q}'(t)A'\underline{x}$ $= \underline{y}'(t)\Lambda[X'A']\underline{x}$ [$\underline{q}'(t)=\underline{p}'(t)A$:Proc. 2]	$y(t) = \underline{b}'(t)Q\underline{x}$ $= [b'(t)Y]\Lambda X'\underline{x}$ [$Q=\langle\underline{b}(\xi),\underline{p}'(\xi)\rangle$:Proc. 1]	$y(t) = \underline{b}'(t)QA'\underline{x}$ $= [\underline{b}'(t)Y]\Lambda[X'A']\underline{x}$ [$Q=\langle\underline{b}(\xi),\underline{p}'(\xi)\rangle A$:Proc. 1]
C-D	$\underline{y} = \langle\underline{p}(\xi),x(\xi)\rangle$ $= Y\Lambda\langle\underline{x}(\xi),x(\xi)\rangle$ [$\underline{p}(t)$: Procedure 3]	$\underline{y} = Q\langle\underline{a}(\xi),x(\xi)\rangle$ $= Y\Lambda\langle X'\underline{a}(\xi),x(\xi)\rangle$ [$Q=\langle\underline{p}(\xi),\underline{a}'(\xi)\rangle$:Proc. 1]	$\underline{y} = B\langle\underline{q}(\xi),x(\xi)\rangle$ $= [BY]\Lambda\langle\underline{x}(\xi),x(\xi)\rangle$ [$\underline{q}(\xi)=B'\underline{p}(\xi)$:Proc. 3]	$\underline{y} = BQ\langle\underline{a}(\xi),x(\xi)\rangle$ $= [BY]\Lambda\langle X'\underline{a}(\xi),x(\xi)\rangle$ [$Q=B'\langle\underline{p}(\xi),\underline{a}'(\xi)\rangle$:Proc. 1]
D-D	$\underline{y} = P\underline{x}$ $= Y\Lambda X'\underline{x}$ [P: Procedure 1]	$\underline{y} = QA'\underline{x}$ $= Y\Lambda[X'A']\underline{x}$ [Q=PA:Procedure 1]	$\underline{y} = BQ\underline{x}$ $= [BY]\Lambda X'\underline{x}$ [Q=B'P:Procedure 1]	$\underline{y} = BQA'\underline{x}$ $= [BY]\Lambda[X'A']\underline{x}$ [Q=B'PA:Procedure 1]
C-C	$y(t) = \langle p(t,\xi),x(\xi)\rangle$ No finite decomposition	$y(t) = \underline{q}'(t)\langle\underline{a}(\xi),x(\xi)\rangle$ $= \underline{y}'(t)\Lambda\langle X'\underline{a}(\xi),x(\xi)\rangle$ [$\underline{q}'(t)=\langle p(t,\xi),\underline{a}'(\xi)\rangle$: Proc. 2]	$y(t) = \underline{b}'(t)\langle\underline{q}(\xi),x(\xi)\rangle$ $=[b'(t)Y]\Lambda\langle\underline{x}(\xi),x(\xi)\rangle$ [$\underline{q}(\xi)=\langle\underline{b}(t),p(t,\xi)\rangle$: Proc. 3]	$y(t) = \underline{b}'(t)Q\langle\underline{a}(\xi),x(\xi)\rangle$ $=[\underline{b}'(t)Y]\Lambda\langle X'\underline{a}(\xi),x(\xi)\rangle$ [$Q=\langle\langle\underline{b}(t),p(t,\xi)\rangle,\underline{a}'(\xi)\rangle$: Proc. 1] [Q: Procedure 1]

First line in each box is the operator representation. Second line is the corresponding decomposition. Third line tells what to decompose and how. If the quantity to be decomposed is $\underline{q}(t)$ or Q, replace $\underline{p}(t)$ by $\underline{q}(t)$ or P by Q in the following procedures:

Procedure 1: Assume P is of rank M. Form W = P'P. Obtain eigenvalue decomposition W = ZHZ', where H is an MxM matrix of the M nonzero eigenvalues of W, and Z is an NxM matrix of the corresponding eigenvectors of W, normalized to unity and taken in the same order as the corresponding eigenvalues in H. Let Λ be the MxM diagonal matrix each entry of which is the positive square root of the corresponding entry of H. Let X = Z and $Y = PX\Lambda^{-1}$.

Procedure 2: Form $W = \langle\underline{p}(t),\underline{p}'(t)\rangle$. Decompose W as in Procedure 1. Set X = Z and $\underline{y}(t) = \Lambda^{-1}X'\underline{p}(t)$.

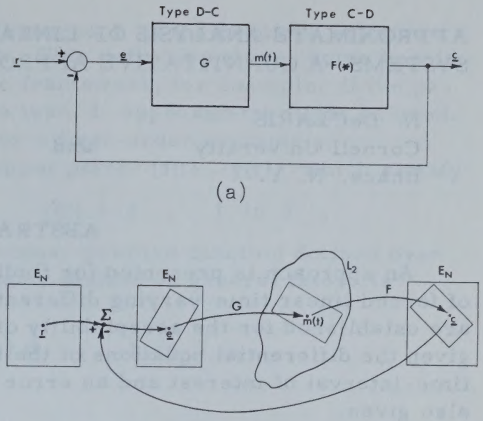
Procedure 3: Form W as in Procedure 2 and decompose as in Procedure 1. Set Y = Z and $\underline{x}(t) = \Lambda^{-1}Y'\underline{p}(t)$.


Type D-C
Type C-C
x
F
y(t)
G(*)
z(t)
y(t) = F(x)
z(t)=G(y(t))=G(F(x))
E_N
F
x
F(E_N)
L_2
y
G
G(F(E_N))
z
G(L_2)

(a)

(b)

Fig. 1a) A typical situation in which only some subspace of the input space of G need be considered. b) The same situation in terms of operators and function spaces.


Type D-C
Type C-D
r
+
−
e
G
m(t)
F(*)
c
E_N
E_N
L_2
E_N
r
Σ
+
−
e
G
m(t)
F
c

(a)

(b)

Fig. 2 Example 1: A feedback control system of Type D-D. a) Block diagram representation. b) Operator-function space representation.

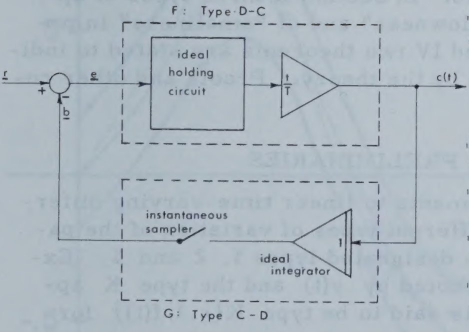
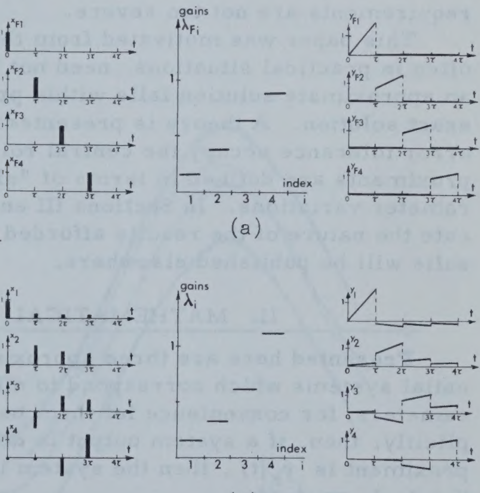

F: Type D-C
r
+
−
b
e
ideal holding circuit
1/T
c(t)
instantaneous sampler
ideal integrator
G: Type C-D

Fig. 3 Example 2: A feedback control system of Type D-C.



(a)

(b)

Fig. 4 Input basis vectors, generalized frequency response, and output basis functions for Example 2 with N = 4 and $\tau = 1$. a) Open-loop system (F only). b) Closed-loop system.

APPROXIMATE ANALYSIS OF LINEAR TIME-VARYING SYSTEMS—A QUANTITATIVE APPROACH*

N. DeCLARIS
Cornell University
Ithaca, N. Y.

and

H. VAN LANDINGHAM
Virginia Polytechnic Institute
Blacksburg, Va.

ABSTRACT

An approach is presented for finding acceptable approximate behavior of forced linear time-varying differential systems. Sufficient conditions are established for the acceptability of convenient approximate analysis given the differential equations of the system, the forcing function, the time-interval of interest and an error tolerance. A simple example is also given.

I. INTRODUCTION

Linear, lumped, finite and time-invariant systems have been the subject of much effort; however, if the last restriction above is removed, thereby allowing time-dependent parameters within the system, exact solutions are virtually impossible to obtain.[1] Thus at the present time, the analysis of linear systems with time-varying parameters rests most often on numerical methods and/or digital computer programs. Such approaches, in general, suffer from two system-theoretic drawbacks: a) the structure of the original model is obscured quite early in the analysis, and b) greater effort is required than is intuitively justifiable for cases where "accuracy" requirements are not too severe.

This paper was motivated from the fact that the analysis of a system often in practical situations, need not be exact, provided it is known that an approximate solution falls within prescribed error tolerance from the exact solution. A theory is presented in which the model structure and error tolerance occupy the central role. In Section II three types of approximants are defined in terms of "slowness" and of "smallness" in parameter variations. In Sections III and IV two theorems are stated to indicate the nature of the results afforded by the theory. Proofs and other results will be published elsewhere.

II. MATHEMATICAL PRELIMINARIES

Presented here are three approximants to linear time-varying differential systems which correspond to different types of variation of the parameters; for convenience let them be designated types 1, 2 and 3. Explicitly, then, if a system output is denoted by $y(t)$ and the type K approximant is $y_k(t)$, then the system is said to be type $K\{\epsilon, I, f(t)\}$ for $K = 1, 2,$ or 3 if

$$\|y_k(t) - y(t)\| = \sup_{t \text{ in } I} |y_k(t) - y(t)| \leq \epsilon \quad ,$$

---

* This work was supported by the U. S. Air Force under Contract AF33(615)-1274.

where ϵ is the prescribed error tolerance, I is the required solution interval and f(t) is the system input. This definition places some heretofore intuitive ideas into a quantitative framework; for example, if the parameter variations occur "slowly," a type 1 approximant might be used. These ideas may well be illustrated by a first-order example.

Let S be a system with input-output pairs $\{f(t)\,,\ y(t)\}$ which satisfy

$$S:\ [D + a(t)]\, y(t) = f(t) \quad , \quad y(0) = 0 \quad , \quad t \text{ in } I \quad ,$$

where $D = d/dt$ and a(t) is a continuous, positive function defined over $I = (0, T)$; then the type 1 approximate, defined in general later, is

$$y_1(t) = \int_0^t \exp\{a(t)\,(t-u)\}\, f(u)\, du \quad , \quad t \text{ in } I \quad ,$$

whereas the exact solution is

$$y(t) = \int_0^t \exp\{p(t) - p(u)\}\, f(u)\, du \quad , \quad t \text{ in } I \quad ,$$

where

$$p(t) = \int_0^t a(v)\, dv \quad .$$

The following graphical presentation, Figure 1, illustrates the definition of type 1 $\{\epsilon, I, f(t)\}$ for the case where

$$a(t) = 1 + \sin(t/10) \quad ,$$

$$f(t) = \sin t \quad \text{and} \quad T = 10$$

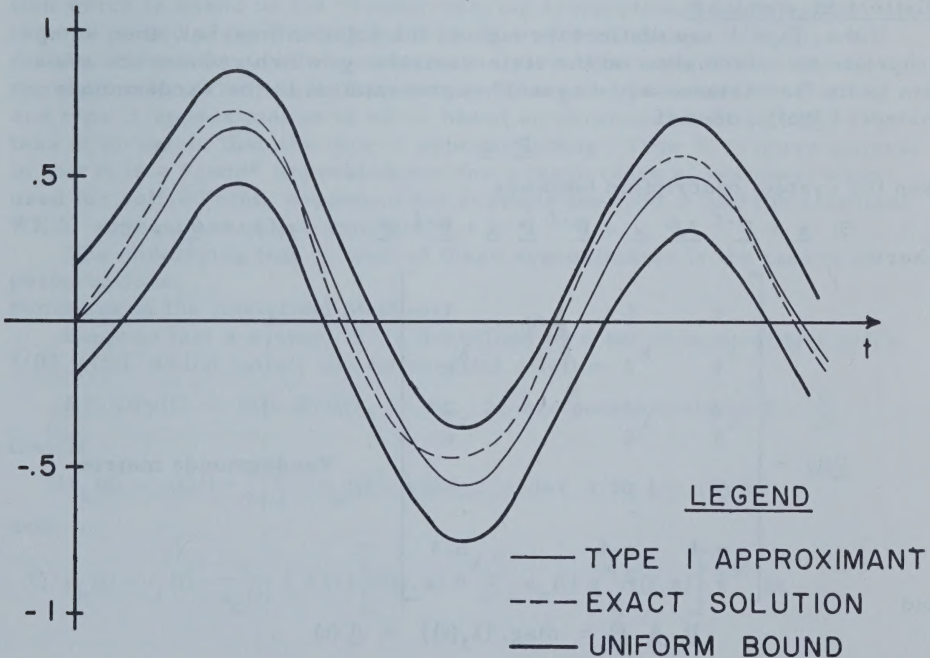


FIGURE 1. An Example.

The uniform bound shown in the graph represents a "tube" around the approximant which is guaranteed to contain the exact solution.

These concepts involving first-order scalar systems become very intractable when being extended to general first-order vector systems. As a simplifying assumption, only systems which are locally passive, i. e. which represent passive systems for each fixed value of time in the solution interval, will be considered.

Let the system be described as follows:

$$S: [D^n + a_{n-1}(t) D^{n-1} + \dots + a_0(t)]\, y(t) = f(t) \quad ,$$

$$y(0) = Dy(0) = \dots = D^{n-1} y(0) \quad \text{for} \quad t \text{ in } I \quad .$$

It is convenient to work from the canonical vector first-order characterization

$$S: D\underline{y}(t) = \underline{A}(t)\ \underline{y}(t) + \underline{F}(t) \quad , \quad \underline{y}(0) = \underline{0} \quad , \quad t \text{ in } I$$

where

$$\underline{y}(t) = \begin{bmatrix} y(t) \\ Dy(t) \\ \vdots \\ D^{n-1} y(t) \end{bmatrix}, \quad \underline{A}(t) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-1} \end{bmatrix}, \quad \underline{F}(t) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ f(t) \end{bmatrix}$$

If it were possible to diagonalize this system, the problem would be reduced immediately to the first-order scalar situation, but since no general method which accomplishes this is presently known, the theory here concerns itself with the instantaneous eigenvalues, $\{\lambda_i(t)\}$, of the system.

Distinct Eigenvalues:

If the $\{\lambda_i(t)\}$ are distinct throughout the solution interval, then an appropriate transformation on the state-variable $\underline{y}$ which reduces the system to its "instantaneously diagonal" representation is the Vandermonde matrix $[\underline{P}(t)]$, thus if

$$\underline{y} = \underline{P}\ \underline{x}$$

then the system description becomes

$$S: \underline{x} = \underline{P}^{-1}\, \underline{AP}\ \underline{x} - \underline{P}^{-1}\, \dot{\underline{P}}\ \underline{x} + \underline{P}^{-1}\ \underline{F} \quad , \quad \underline{x}(0) = \underline{0} \quad ,$$

where

$$\underline{P}(t) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & & \lambda_n^2 \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & & \lambda_n^{n-1} \end{bmatrix}, \quad \text{Vandermonde matrix}$$

and

$$\underline{P}\ \underline{A}\ \underline{P} = \text{diag.}\,\{\lambda_i(t)\} = \underline{\Lambda}\,(t) \quad .$$

The restriction on S that $[\underline{P}^{-1}\,\dot{\underline{P}}]$ be small compared with $[\Lambda]$ implies that the variations in S occur "slowly." The type 1 approximant is, in fact, a model which neglects the effects of the time derivatives of the system function $H(s,t)$.[2] The mathematical approach to the bounding problem is to use perturbation techniques; i. e., to consider the matrix $[\underline{P}^{-1}\,\dot{\underline{P}}]$ as a perturbation on the diagonal system and truncate the resulting series expression of the exact solution. Thus, the a priori error bound must be divided into two parts to accommodate the error bound of the difference between the exact and approximate solutions which results from two bounding procedures: 1) the norm of the difference between the exact and the instantaneously diagonal system and 2) the maximum first-order error of the diagonal system.

Nondistinct Eigenvalues:

If at some time t_1 in the solution interval two or more of the instantaneous eigenvalues become equal, then the perturbing matrix $[\underline{P}^{-1}\,\dot{\underline{P}}]$ becomes unbounded as t approaches t_1. For this reason yet another error is incurred by using a constant-system model for intervals about such points. The length of these intervals is a function of the minimum distance, δ, allowed between eigenvalues; the set

$$\Sigma_\delta = \{t \in I : \; |\lambda_k(t) - \lambda_j(t)| \leq \delta \;, \quad j \neq k\}$$

is called the set of turning intervals.

The results of this paper show, however, that for special classes of driving functions, the overall error bound, the sum of each bounding process, can be arbitrarily small.

The remarks made so far have been mainly for the type 1 approximation which is based on the "slowly-varying assumption," namely that during one period of the driving function, the system change is negligible, that is, a restriction of slowness on the parameter variation. The type 2 approximation is similarly based on smallness of parameter variation, and type 3 approximation is again based on slowness of variation but to less of an extent than the type 1 approximation. Type 3 is more general in that it is a "good" approximation for a large class of systems; when used for self-adjoint, second-order systems the type 3 and the classical WKBJ approximation are equivalent.

The underlying link to each of these approximants is the theory of perturbations.

Summary of the Analytical Methods

Suppose that a system S is described by a set of input-output pairs $\{f(t), y(t)\}$ which satisfy the differential relation

$$N(t,D)\,y(t) = M(t,D)\,f(t) \quad ; \quad N,\; M \text{ are polynomials in } D = \frac{d}{dt}$$

then if

$$||y_k(t) - y(t)|| = \sup_{t \in I} |y_k(t) - y(t)| \leq \epsilon \quad \text{for } t \text{ in } I = [0,T]$$

and:

$$1)\quad y_k(t) = y_1(t) = \frac{1}{a_0(t)} \int F(s) \left[M(t,s) - \sum_{k=1}^{n} a_k(t)\, s^k H(t,s) \right] e^{st}\, ds$$

then S is type 1 (ϵ, I, f) ;

$$2)\quad y_k = y_2(t) = \int F(s) \left[\frac{M(t,s)}{\bar{\alpha}_0(s)} - \frac{1}{\bar{\alpha}_0(s)} \sum_{i=1}^{n} \bar{\alpha}_i(s)\, H(s,t) \right] e^{st}\, ds$$

then S is type 2 (ϵ, I, f) ;

3) lastly, for a state-space description of S :

$$\dot{\underline{y}}(t) = \underline{A}(t)\, \underline{y}(t) + \underline{f}(t) \quad ; \quad \underline{y}(t_0) = \underline{y}_0 \, , \quad t_0 \in I$$

$$\underline{y}(t) = \Phi_A(t, t_0)\, \underline{y}_0 + \int_{t_0}^{t} \Phi_A(t, \lambda)\, \underline{f}(\lambda)\, d\lambda \quad , \text{ and}$$

$$\underline{y}_3(t) = \theta_A(t, t_0)\, \underline{y}_0 + \int_{t_0}^{t} \theta_A(t, \lambda)\, \underline{f}(\lambda)\, d\lambda \;\; \forall \; t \in I$$

then S is a type 3 (ϵ, I, f) if

$$\theta(t, \tau) = \underline{P}(t)\, \underline{E}(t, \tau) \left[\underline{1} - \int_{\tau}^{t} \underline{E}^{-1}(s, \tau)\, \underline{P}^{-1}(s)\, \dot{\underline{P}}(s)\, \underline{E}(s, \tau)\, ds \right] \underline{P}^{-1}(\tau)$$

where $\underline{P}(t)$ satisfies $\underline{P}^{-1} \underline{A}\, \underline{P} = \underline{\Lambda}(t)$, diagonal matrix of the eigenvalues of the matrix $\underline{A}(t)$ and

$$\underline{E}(t, \tau) = \exp \int_{\tau}^{t} \underline{\Lambda}(s)\, ds \quad .$$

Notation:

$$N(t, D) = D^n + a_{n-1}(t)\, D^{n-1} + \ldots + a_0(t)$$

$$F(s) = \text{Laplace transform of } f(t)$$

$$\alpha_k(t \;\, s) = \frac{1}{k!} \frac{\partial^k}{\partial s^k} N(t, s) \quad , \qquad k = 0, 1, \ldots, n$$

$$\bar{\alpha}_k(s) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \alpha_k(s, t)\, dt \quad , \qquad k = 0, 1, \ldots, n \quad .$$

Thus, each type of approximant involves approximating one of the system functions, i. e. in types 1 and 2 the frequency-response function is approximated, and in the type 3 the impulsive response function for the particular configuration is approximated.

III. MAIN RESULTS

Thus far the three different approximants have only been motivated and outlined. In this section is presented a theorem which is representative of the results obtainable from the methods previously discussed.

Theorem 1 :

The zero-state, locally-passive system S with input-output pairs described by $\{f(t) \, , \; y(t)\}$

$S : [D^2 + a_1(t)\, D + a_0(t)]\; y(t) = f(t)$, where a_1 and a_0 are continuous functions, is type 1 (ϵ, I, f) for $f(t)$ in the set P_0, continuous and piecewise smooth providing that

$$\omega_0 \ge \frac{\pi \nu}{6 \epsilon}\left\{ \bar{T}\, e^{\underline{\sigma}\,\bar{T}}\, \frac{2}{\underline{\delta}}\, m\left[\frac{1}{\underline{\delta}}\, e^{\underline{\sigma}\,\bar{T}}(1+\bar{T}\,\bar{\sigma})\right] + \frac{1}{\underline{\delta}}\left(1+m^{\frac{1}{2}}\, e^{-\frac{1}{2}}\right)\right.$$

$$\left. + \sum_{k=1}^{r}\left[M_0^{(k)}\left(1+\Delta_k \ln M_0^{(k)}\right)\left(e^{M_0^{(k)}\Delta_k}-1\right) + \Delta_k(1+\Delta_k)\left|e^{\Delta_k \sigma_k}-1\right|\right]\right\}$$

where

$$P_0 = \left\{ f(t) : f(t) = f\left(t+\frac{2\pi}{\omega_0}\right) \;,\; \int_0^{2\pi/\omega_0} f(x)\, dx = 0 \;,\; ||f|| = 1 \right\}$$

ν = variation of $f(t)$ over $\left[0, \frac{2\pi}{\omega_0}\right]$

$\bar{T}$ = length of the maximum interval of the complement of Σ_δ

$\underline{\sigma} = \inf\limits_{k, t \in I} \operatorname{Re}\{\lambda_k(t)\}$

$\bar{\sigma} = \sup\limits_{k, t \in I} \operatorname{Re}\{\lambda_k(t)\}$

$\underline{\delta} = \inf\limits_{t \notin \Sigma_\delta} |\lambda_1(t) - \lambda_2(t)|$

$m = \sup\limits_{k, t \in I} \operatorname{Re}\{\dot{\lambda}_k(t)\}$

r = number of turning intervals

$M_0^{(k)} = \max\left\{ e^{\Delta_k}\; e^{\Delta_k(a_0 + a_1)} \right\}$

$a_k = ||a_k(t)||$, $k = 0, 1$

Δ_k = length of the k^{th} interval of Σ_δ

$\sigma_k = \operatorname{Re}\{\lambda(t_k)\}$

t_k = center of the k^{th} turning interval

Lack of space prohibits the inclusion of theorem proofs; for these, the reader is referred to forthcoming publication.[3]

The exact statements for higher order systems involve even more notational difficulty and will be published elsewhere.

IV. APPROXIMATIONS AND SYSTEM EQUIVALENCE

The bounding techniques used on a specific system actually generate classes of systems. The following theorem is pertinent.

Theorem 2:

Any system which is uni-kinematically similar to a system S_0 having a uniform bound for a type K approximant, $k = 1, 2, 3$, also satisfies these same bounds. A uni-kinematically similarity transformation is used to denote a Lyapunov transformation having a unit norm on the state-space representation of the system.

It is clear that the constant parameter systems are the "kernel" of each of the classes generated via Theorem 2 since any constant system satifies the definition of each approximant to an arbitrary degree of accuracy. With Theorem 2, then, the techniques indicated here generate a one-parameter family of classes of systems which satisfy the given type of requirement and class of inputs.

V. REFERENCES

1. L. A. Zadeh and C. A. Desoer, Linear System Theory, New York: McGraw-Hill, 1963.
2. J. B. Cruz and M. E. Van Valkenburg, "Synthesis of Time-Varying Linear Systems," Symposium on Active Networks, Polytechnic Institute of Brooklyn, 1960.
3. H. F. Van Landingham, Ph. D. Thesis, and Research Technical Report, Cornell University, to be published.

ESTIMATION OF THE IMPULSE RESPONSE OF TIME-VARYING LINEAR SYSTEMS

WILLIAM M. HAMMOND
Bradley University
Peoria, Illinois

GEORGE R. COOPER
Purdue University
Lafayette, Indiana

ABSTRACT

When crosscorrelation is used to experimentally determine the impulse response of a time-varying linear system, the correlator output cannot be smoothed for an arbitrarily long period of time. If the optimum weighting function of the output smoothing filter is taken to be that which minimizes the mean squared error in the estimate of the system impulse response for all time, then it is time varying and can be specified in terms of a Fredholm integral equation of the first kind with a non-separable kernel. Since a closed-form solution of this equation is not available in general, an approximate solution based on matched filter concepts, and leading to a suboptimum weighting function, is proposed. The resulting mean squared estimation error is evaluated and compared with the optimum case.

INTRODUCTION

A well-known method of experimentally determining the impulse response of a linear system is to crosscorrelate the input and output of the system when the input is a sample function from a wideband stationary random process.[1] When the system being tested is time invariant, the resultant estimate of the impulse response is unbiased and can be made to have an arbitrarily small variance by increasing the averaging time of the correlator output. This is not true, however, if the system is time varying because increasing the averaging time beyond some optimum value will usually lead to an increased bias in the estimate even though the variance of the estimate becomes smaller. Furthermore, a uniform weighting of the data being averaged is no longer the best strategy. It is of interest, therefore, to determine the weighting which is best according to some criterion of goodness and to compare the estimates obtained in this way with estimates obtained from other possible weightings.

The approach to this problem considered here is to assume that the averaging of the correlator output is accomplished by means of a linear, physically-realizable filter which may also be time varying. The optimum filter is taken to be that which minimizes the mean squared error of the estimate at every instant of time and the impulse response of this filter can be

---

This paper is based on a dissertation submitted by the first author to the Faculty of Purdue University in partial fulfillment of the requirements for the Ph.D. degree. The work was supported by the National Science Foundation under Grant No. GK-189.

expressed in terms of an integral equation which cannot be solved readily in closed form. The contribution of the present work is an approximate solution to this integral equation and an evaluation of the estimation errors from the corresponding suboptimum filter.

THE OPTIMUM SMOOTHING FILTER

A block diagram of the method used for estimating the impulse response of a linear system is shown in Fig. 1. The system being tested has an impulse response of $h_m(t,\xi)$ and is excited by a sample function, $y(t)$, from a wideband stationary, Gaussian, random process. For convenience in discussion, this process will be assumed to be white noise, but the use of a more realistic spectrum will not alter the major results in any significant way. The output of the system and a delayed version of the input are multiplied and averaged in the smoothing filter having an impulse response of $h(t,\xi)$. The determination of $h(t,\xi)$ is the central problem in this discussion.

The multiplier output, $z(t)$, is given simply by

$$z(t) = \int_{-\infty}^{t} y(t-\tau)\, y(\xi)\, h_m(t,\xi)d\xi \tag{1}$$

and the expected value of this time function is

$$E[z(t)] = \overline{z}(t) = \int_{-\infty}^{t} R_y(\xi,\ \xi-t+\tau)h_m(t,\xi)d\xi \tag{2}$$

where the autocorrelation function of $y(t)$ is defined in general by

$$R_y(t,\ \lambda) = E[y(t)\ y(t-\lambda)]$$

Under the assumption that $y(t)$ is a sample function of white noise with a spectral density of Y_o, (2) becomes

$$\overline{z}(t) = Y_o\, h_m(t,\ t-\tau) \tag{3}$$

Hence, the purpose of the smoothing filter is to produce an estimate of the mean value of $z(t)$.

For practical reasons the smoothing filter is assumed to have its input signal, $z(t)$, available over a finite time interval of T seconds only. This restriction is necessary for digital computations and can be relaxed for analytical purposes simply by letting $T \to +\infty$. Thus, the estimate of the system impulse response can be expressed as

$$\hat{h}_m(t,\ t-\tau) = \int_{t-T}^{t} z(\xi)\, h(t,\xi)d\xi \tag{4}$$

and the error in this estimate as

$$\begin{aligned} e(t,\tau) &= h_m(t,\ t-\tau) - \hat{h}_m(t,\ t-\tau) \\ &= h_m(t,\ t-\tau) - \int_{t-T}^{t} z(\xi)h(t,\xi)d\xi \end{aligned} \tag{5}$$

The optimum form of $h(t,\xi)$ is that which minimizes the mean squared value of the error, $E[e^2(t,\tau)]$. It can be shown by straightforward calculus of variations techniques that the desired $h(t,\xi)$ must satisfy an integral equation of the form[2,3]

$$\int_{t-T}^{t} h(t,\lambda)\,\mu_z^2(\xi,\lambda,\tau)\,d\lambda = Y_o^2 h_m(t,t-\tau)h_m(\xi,\xi-\tau), \qquad t-T < \xi < t \tag{6}$$

where

$$\begin{aligned}\mu_z^2(\xi,\lambda,\tau) &= E[z(\xi)\;z(\lambda)]\\ &= E[x(\xi)\;y(\xi-\tau)\;x(\lambda)\;y(\lambda-\tau)]\end{aligned}$$

is a fourth product moment of x and y and is usually not separable into the product of a function of ξ only and a function of λ only. Hence, this is a Fredholm integral of the first kind with a non-separable kernal. Closed-form solutions for this type of integral equation are not known in general.

THE SUBOPTIMUM SMOOTHING FILTER

Since exact solutions to (6) are not available, it is of interest to seek approximate ones. For this purpose it is convenient to represent the input to the smoothing filter as the desired component and noise. That is, let

$$z(t) = Y_o h_m(t,t-\tau) + n(t) \tag{7}$$

Since the desired component is just $\overline{z}(t)$, it is clear that n(t) must have zero mean value for all t. Furthermore, if any significant smoothing is going to be possible, then n(t) must have a bandwidth which is large compared to the bandwidth of the smoothing filter. Under these circumstances a reasonable approximation is to replace n(t) by a sample function of white noise having a spectral density of N_o, and introduce this assumption into the integral equation (6).

When the above procedure is carried out, the integral equation for the suboptimum smoothing filter becomes

$$\begin{aligned}N_o h_s(t,\xi) + Y_o^2 \int_{t-T}^{t} h_s(t,\lambda)h_m(\xi,\xi-\tau)h_m(\lambda,\lambda-\tau)\,d\lambda\\ = Y_o^2\, h_m(t,t-\tau)h_m(\xi,\xi-\tau),\\ t-T < \xi < t\end{aligned} \tag{8}$$

where $h_s(t,\xi)$ is the impulse response for the suboptimum smoothing filter. This integral equation is a Fredholm equation of the second kind with a separable kernal for which there are well-known methods of solution. The resulting solution is

$$h_s(t,\xi) = \frac{Y_o^2 h_m(t,t-\tau)\;h_m(\xi,\xi-\tau)}{N_o + Y_o^2 \int_{t-T}^{t} h_m^2(\lambda,\lambda-\tau)\,d\lambda}, \quad t-T < \xi < t \tag{9}$$

There still remains the question of selecting an appropriate value for N_o. In accordance with our earlier philosophy of minimizing the mean squared error in the estimate,

$\hat{h}_m(t,\xi)$, a reasonable criterion is to permit N_o to be the function of time which does minimize this error when the smoothing filter has the form specified by (9). Straightforward substitution into a general expression for mean squared error, and subsequent minimization with respect to N_o, leads immediately to

$$N_o = \frac{\int_{t-T}^{t} d\lambda \int_{t-T}^{t} [h_m(\xi,\ \xi-\tau)\ h_m(\lambda,\ \lambda-\tau)\mu_z{}^2(\xi,\ \lambda,\ \tau)}{\int_{t-T}^{t} h_m{}^2(\lambda,\ \lambda-\tau)d\lambda}$$

$$- \frac{Y_o{}^2\ h_m{}^2(\xi,\ \xi-\tau)\ h_m{}^2(\lambda,\ \lambda-\tau)]d\lambda}{\int_{t-T}^{t} h_m{}^2(\lambda,\ \lambda-\tau)d\lambda} \tag{10}$$

Note that this is a function of time, t.

COMPARISON OF MEAN SQUARED ERRORS

In order to compare the optimum and suboptimum smoothing filters, it necessary to calculate the mean squared errors associated with the estimates produced by each. In the case of the suboptimum filter this is easily accomplished by inserting (9) and (10) into the general expression for mean squared error and simplifying. The result is

$$E[e_s{}^2(t,\tau)]_{min} = \frac{N_o h_m{}^2(t,\ t-\tau)}{N_o + Y_o{}^2 \int_{t-T}^{t} h_m^2(\lambda,\ \lambda-\tau)d\lambda} \tag{11}$$

where N_o is that specified in (10). It is of interest to note the equivalence (except for a scale factor) of (11) and (9) when $\xi = t$. This is completely analogous to a similar result for the matched filter which maximizes signal-to-noise ratio in white noise at any time t.[4]

Since the exact form of the optimum smoothing filter cannot be obtained, it is necessary to resort to a discrete approximation in order to evaluate the minimum mean squared error in this case. This is done by sampling z(t) at equal time intervals and by representing the filter by a set of samples of its impulse response. The integral equation (6) can then be represented in matrix form and the optimum impulse-response vector solved for explicitly. Substitution into the error expression leads to a quadratic form which, when evaluated, yields the minimum mean squared error for the discrete optimum smoothing filter. If the sampling interval is sufficiently small, this mean squared error should be a good approximation to the true minimum mean squared error for the continuous optimum smoothing filter. Since the use of the discrete filter is simply a computational technique, there is no need to discuss the formulation of this procedure in detail.

Because different procedures are necessary in order to evaluate the mean squared errors for the optimum and suboptimum cases, it is not possible to make a general comparison of these two filters on an analytical basis. Specific numerical comparisons can be made, however, for any given example. Such an example is discussed in the following section.

TIME-VARYING, FIRST-ORDER SYSTEM

For purposes of illustration, consider a first-order system with a time-varying gain. Specifically, the impulse response to be estimated is assumed to have the form

$$h_m(t,\xi) = \varepsilon^{-\beta\xi}\,\varepsilon^{-\alpha(t-\xi)}, \quad t \geq \xi, \quad \alpha > \beta \qquad (12)$$
$$= 0 \qquad\qquad , \quad t < \xi$$

The impulse response of the corresponding suboptimum smoothing filter, as given by (9), has been computed for particular values of the parameters and is shown in Fig. 2 along with the system impulse response for the same parameters. The matched filter analogy is clearly evident here.

The mean squared errors resulting from the suboptimum smoothing filter and the corresponding optimum smoothing filter (as computed from the discrete approximation) are shown in Figs. 3, 4, and 5 for several different values of the parameters. In all cases the spectral density Y_o of the system input, y(t), has been taken as unity. The close comparison between the values for the optimum and suboptimum cases indicates that the approximate solution yielded by suboptimum approach must be very close to the true solution.

CONCLUSIONS

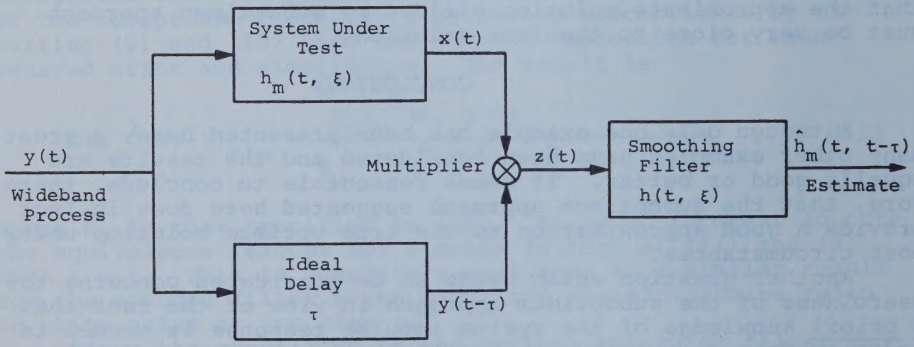
Although only one example has been presented here, a great many other examples have been considered and the results are equally good or better. It seems reasonable to conclude, therefore, that the suboptimum approach suggested here does in fact provide a good approximation to the true optimum solution under most circumstances.

Another question which needs to be considered concerns the usefulness of the suboptimum approach in view of the fact that *a priori* knowledge of the system impulse response is needed in order to determine either the optimum or suboptimum smoothing filter. This, of course, is the characteristic weakness of all conditional estimation methods and is usually rationalized on the basis that it gives the engineer additional insight into the design of more practical filters when a limited amount of *a priori* information is available. In the present case, such insight must come from the suboptimum case since explicit, continuous solutions cannot be obtained for the true optimum.

A final question concerns the need for a suboptimum approach in situations which must be relegated to the computer since a discrete approximation is available for the optimum case. The difficulty here is that the number of samples needed for a good approximation may be so large (for an oscillatory system, for example) that the computer solution for the discrete optimum filter becomes unmanageable. If the number of samples is severely limited by computer storage, then the suboptimum approach (which has no such limitation) will provide a better approximation than the discrete optimum approach.

REFERENCES

1. Lee, Y. W., Statistical Theory of Communication, John Wiley and Sons, New York, 1960. Chap. 13.
2. Hammond, W. M., "Real-Time Estimation of Time-Varying Correlation Functions," Ph.D. Thesis, Purdue University, January 1967.
3. Berndt, H., and G. R. Cooper, "Estimates of Correlation Functions of Nonstationary Random Processes", IEEE Trans. on Information Theory, Vol. IT-12, No. 1, January 1966, pp. 70-72.
4. Davenport, W. B., Jr., and W. L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, New York, 1958.



System Under Test
$h_m(t, \xi)$
x(t)
y(t)
Wideband Process
Multiplier
z(t)
Smoothing Filter
$h(t, \xi)$
$\hat{h}_m(t, t-\tau)$
Estimate
Ideal Delay
τ
$y(t-\tau)$

Fig. 1. Block diagram of Estimating Method.

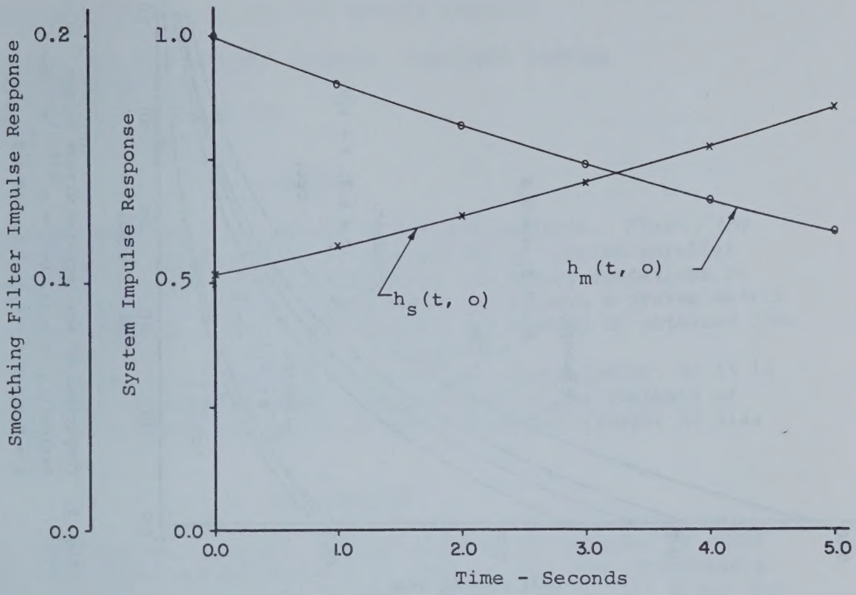


Fig. 2. Impulse Responses of System, $h_m(t, o)$, and Corresponding Suboptimum Smoothing Filter, $h_s(t, o)$.

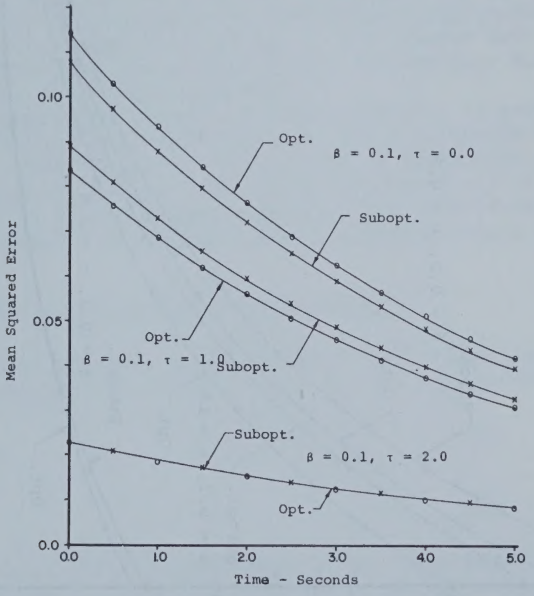


Fig. 3. Comparison of Mean Squared Errors in the Estimation of $h_m(t, t-\tau) = e^{-\beta(t-\tau)}e^{-\tau}$ for $\beta = 0.1$ and $\tau = 0.0, 1.0,$ and 2.0.

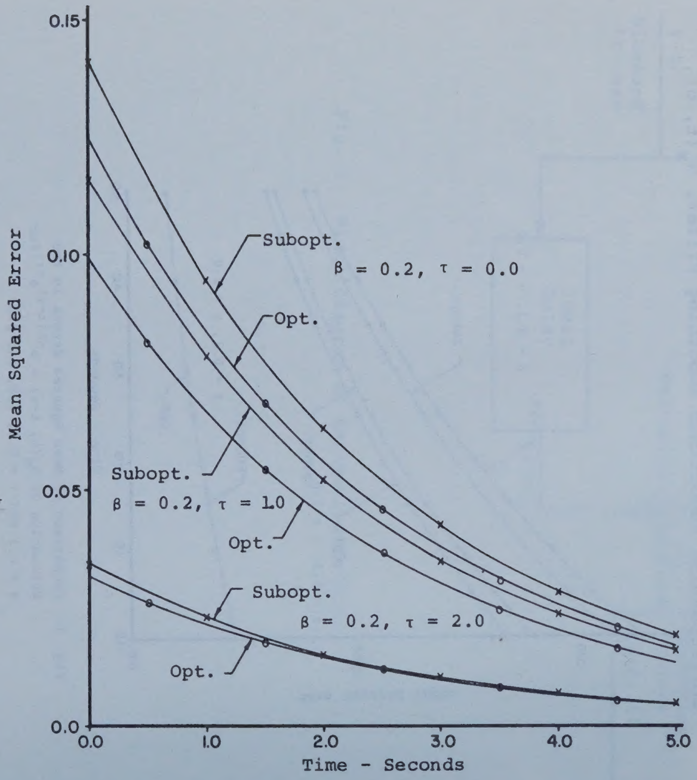


Fig. 4. Comparison of Mean Squared Errors in the Estimation of $h_m(t, t-\tau) = e^{-\beta(t-\tau)}e^{-\tau}$ for $\beta = 0.2$ and $\tau = 0.0$, 1.0, and 2.0.

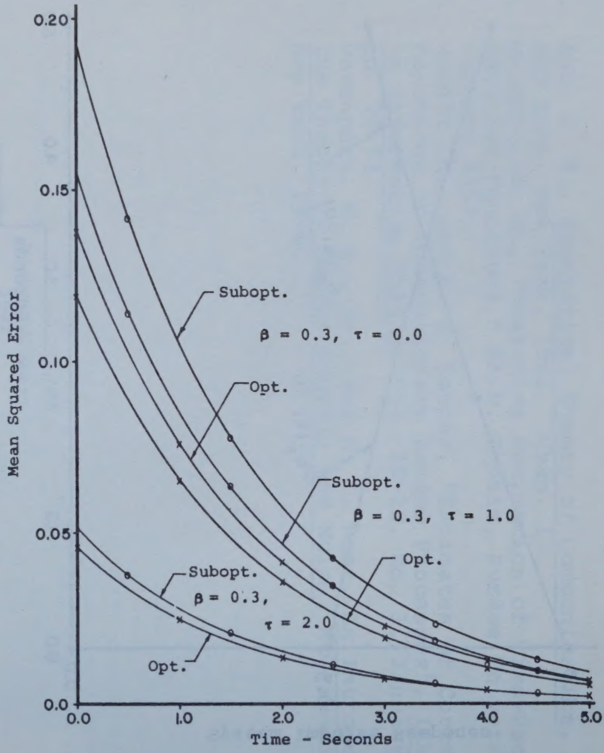


Fig. 5. Comparison of Mean Squared Errors in the Estimation of $h_m(t, t-\tau) = e^{-\beta(t-\tau)}e^{-\tau}$ for $\beta = 0.3$ and $\tau = 0.0$, 1.0, and 2.0.

AN ALGEBRAIC MANIPULATION METHOD FOR NETWORK ANALYSIS

YASUTOSHI ISHIZAKI, TOHORU TSUCHIYA, NOBUYOSHI YOSHIDA
and HITOSHI WATANABE
Nippon Electric Company Ltd.
Kawasaki City, Japan

ABSTRACT

This paper presents a new method of network analysis. First, the concept of series-parallel network is generalized as "series-parallel n-port networks" by defining unary operations and binary operations on port by port of n-port networks. Using these operations, a system matrix of complicated series-parallel n-port network can readily be obtained from given system matrices of rather simple subnetworks.

All procedures of this method are expressed algebraically, so it is very suitable for computer programs. Applications to the analysis of various kinds of networks are discussed and a numerical example is also given.

I INTRODUCTION

Among many network problems, network analysis by computer is still one of the most important problem and in order to save the computational time many attempts have been made by many authors. It is well known that analysis of some special networks such as series-parallel network or ladder network is fairly simple and computer programs for such special network analysis were developed in Japan by authors(1) and in other countries(2). However, this special method is not always applicable to usual practical problems. Average networks with which design engineers often encounter are not so complicated and may be partitioned into subnetworks. If these partioning characters are used effectively, there may be some possibilities to save computation time.

In this paper, a new useful method of network analysis is proposed which is suitable for automatic computation with less computation time compared with ordinary nodal analysis. This method is applicable to newly definied "series-parallel n-port networks" which consist of set of n-ports connected port by port in series and/or parallel. This type of networks is of course no longer general but it has enough generality when n becomes large.

II ELEMENTARY OPERATIONS ON SYSTEM MATRIX

Any linear n-port network can be specified by a set of n independent linear equations with respect to n port-voltages V_j and n port-currents I_j as described below:-(3)

$$[P;Q]\begin{bmatrix}V\\I\end{bmatrix}=0 \tag{1}$$

or

$$M\cdot X=0 \tag{2}$$

where $M=[P;Q]$ is an nx2n matrix of rank n, $X=\begin{bmatrix}V\\I\end{bmatrix}$ is a 2nx1 vector. Matrix M is usually called as "System Matrix" of the n-port network. If a system matrix is non-separable one, then we shall call n-port corresponding to this system matrix a non-separable n-port network.

Now, we shall define following elementary operations of system matrices.

(i) Permutation of Port-Numbers.

Changing the port-number; this corresponds to permutations of columns of system matrix.

(ii) Sign change of a Port.

This corresponds to insertion of ideal transformer of 1 to -1.

(iii) Open Circuiting of a Port.

One port, say i-th port, is open circuited and is no longer regarded as a port. This process consists of following eliminating operations; (1) remove i-th column of nxn submatrix Q and (2) pivotal condensation of i-th column of nxn submatrix with respect to arbitrary non-zero entry of the column. Resulting system matrix is (n-1)x2(n-1) matrix.

(iv) Short Circuit of a Port.

One port, say i-th port, is short circuited and is no longer regarded as a port. This process also corresponds to similar elimination procedure as the above. Resulting system matrix is also (n-1)x2(n-1)matrix.

(v) Series Connection of Two Ports.

Take two ports among n ports and connect them in series as shown in Fig. 1, then resulting network becomes (n-1)-port.

(vi) Parallel Connection of Two Ports.

Connecting two ports in parallel as shown in Fig. 2. This procedure also reduces number of ports by one and resulting network is (n-1)-port.

(vii) Combination of n-port and m-port.

Suppose n-port N_n and m-port N_m and regard both network as an (n+m)-port network as shown in Fig. 3.

Applying these seven types of elementary operations, various networks are formed starting from rather simple subnetworks or, oppositly, complex networks are decomposed into rather simple subnetworks. The system matrix of the final network is easily obtained from system matrices of component subnetworks.

III ALGEBRAIC EXPRESSIONS OF CONNECTIONS OF n-PORTS

For the sake of simplicity, we shall classify elementary operations stated above into two classes, and introduce symbols of operations as explained below.

(A) Unary Operation

It can be stated without loss of generality that elementary operations (iii) and (iv) are applied to non-separable n-port network only. Therefore, these operations can be expressed as

$$\alpha\, k_1k_2 \ldots k_n(N_1)=N_2 \qquad (3)$$

where N_1 is a non-separable n-port network and $\alpha\, k_1k_2 \ldots k_n$ is a symbol of operation in which

$$k_i=\begin{cases} \cdot & \text{non-effect for i-th port} \\ 0 & \text{short-circuit and eliminate i-th port} \\ 1 & \text{open-circuit and eliminate i-th port} \end{cases}$$

Equation (3) symbolizes unary operation of n-port network. For example, Fig. 4 shows 1-port network N_2 derived from 3-port network N_1 by unary operation $\alpha{\cdot}10$. Evidently, this unary operation always reduces number of ports, and an algorithm for finding system matrix of N_2 from that of N_1 is as follows:-

(a) remove columns of P and Q which correspond to $k_i=1$ and 0, respectively

(b) eliminate columns of P and Q which correspond to $k_i=0$ and 1, respectively.

(B) Binary Operation

It is practically useful that elementary operations(v) and (vi) accompanied with operation (ii) are applied between corresponding ports of two different n-ports. In this sense, these operations are concidered as binary operations on two n-ports and can be symbolized as following equation.

$$N_1 \beta_{k_1 k_2 \ldots k_n} N_2 = N_3 \qquad (4)$$

where N_1 and N_2 are n-port networks and $\beta_{k_1 k_2 \ldots k_n}$ is a symbol of a set of elementary operations (ii),(v) and (vi) in which

$$k_i = \begin{cases} . & \text{non-effect} \\ 0 & \text{connect in parallel two i-th ports of } N_1 \text{ and } N_2 \\ 2 & \text{change sign of i-th port of } N_2 \text{ and connect in parallel} \\ & \text{with i-th port of } N_1 \\ 1 & \text{connect in series two i-th ports of } N_1 \text{ and } N_2 \\ 3 & \text{change sign of i-th port of } N_2 \text{ and connect in series} \\ & \text{with i-th port of } N_1 \end{cases}$$

Number of ports of resulting multi-port N_3 in (4), say r, is restricted as

$$n \leqslant r \leqslant 2n-1 \qquad (5)$$

For example, Fig. 5 shows some of resulting networks derived from two 2-ports by applying several operations such as $\beta_{.0}$, $\beta_{.1}$, β_{01}, ..., β_{13}, respectively.

Suppose N_1 and N_2 are specified by following system matrices with nxn submatrices P_1, Q_1, P_2 and Q_2

$$N_1 : [P_1 : Q_1]$$
$$N_2 : [P_2 : Q_2],$$

then an algorithm for this binary operations are summarized as follows:-

If $k_i=0(1)$ $\{k_i=2(3)\}$, $(i=1,\ldots,n)$, then

(a) add {subtract} i-th column of P_2 (Q_2) to i-th column of P_1 (Q_1), and subtract {add} i-th column of Q_1 (P_1) from i-th column of Q_2(P_2),

(b) remove i-th column of P_2 (Q_2),

(c) eliminate i-th column of Q_2 (P_2).

IV SERIES-PARALLEL n-PORT NETWORKS

It is well known that a series-parallel network is such network whose immittance is readily obtained from immittances of the component subnetworks by symple algebraic laws:(4)

$$Z_3=Z_1+Z_2 \quad \text{if } Z_1,\ Z_2 \text{ are connected in series} \qquad (6)$$

$$Y_3=Y_1+Y_2 \quad \text{if } Y_1,\ Y_2 \text{ are connected in parallel} \qquad (7)$$

This property is the simplest case of our binary operations described above. Namely, suppose two 1-ports N_1, N_2 and apply (6), then we have

$$N_1 \beta_k N_2 = N_3 \tag{8}$$

Since sign change of N_1 or N_2 does not change the value of the immittance, β_k is simply reduced to $\beta_.$(non-effect), β_o(parallel connection) or β_1(series connection), from this fact, following definition can be stated.

[Definition 1]
A series-parallel 1-port network with respect to S_1 is the network which is generated by recursive applications of binary operation (4) to elements belonging to S_1, where S_1 is a set of certain types of 1-ports.

As S_1, we can choose any set of 1-ports. For example, when S_1 is a set of branches of L,C or R, the defined series-parallel 1-port is nothing but the ordinary series-parallel network. S_1 can be chosen as a set of switching elements, relays, diodes or distributed line impedances and so forth.

On the similar manner, an extended series-parallel conception can be defined as follows:-

[Definition 2]
A series-parallel n-port network with respect to S_n is the network which is generated by recursive applications of unary operation (3) and binary operation (4) to elements belonging to S_n, where S_n is a set of types of n-ports.

Let n=2 for example, then S_2 may be chosen as a set of ladder networks, nonsymmetrical lattice networks, distributed transmission lines, transistors or transformers. In general case, $n \geqslant 3$, S_n may be a set of hybrid transformers, multi-wire transmission lines, multi-electrode tubes or combinations of subnetworks with less number of ports.

As described before, a binary operation of two 2-ports yields at most 3-port. Therefore, in order to continue the recursive application of binary operation (4) the number of ports of the generated 3-port should be reduced by applying unary operation. For instance,

$$(\text{2-port})\ \beta_{oo}(\text{2-port})=(\text{2-port}) \tag{9}$$

$$\alpha_{\cdot\cdot 1}((\text{2-port})\beta_{.o}(\text{2-port}))=(\text{2-port}) \tag{10}$$

These expressions mean network connections as shown in Fig. 6.

Therefore, system matrix or any type of network parameter of series-parallel n-port can be readily found by simple algebraic expressions which consist of unary operations and binary operations with permutation of port numbers. This is the most important property of "series-parallel n-ports" in both theoretical and practical aspects.

V COMPUTER PROGRAM AND EXAMPLES

If a given network is a series-parallel n-port network with pertinent degree n, then the analysis of this network is simple application of the elementary operations. It can be easily realized that this approach saves the time of computation considerably compared with the ordinary node or loop analysis. Moreover, this method is very suitable for the automatic computation and it has been applied to many fields for several years.

In our program, S_1, S_2 and S_3 contain following types of networks.

S_1 : branches such as L,R and C,
S_2 : ladder networks, nonsymmetrical lattices, uniform or dispersive transmission lines and non-bilateral 2-ports having Z or Y matrix,
S_3 : coupled three wire or hybrid transformers.

Some typical examples how to express networks algebraically are illustrated in Fig. 7.

Fig. 8 is an example of computation. The results of network analysis are represented in terms of attenuation characteristics and phase characteristics. Immittance characteristics can also be obtained if required.

VI CONCLUSION

Elementary operations on ports of n-port are considered and a new concept "series-parallel n-port network" is proposed whose connection status can be described by means of algebraic expression. Network analysis of series-parallel n-port networks is always carried out from system matrices of simpler subnetworks in terms of algebraic treatments. This method is effective not only to lumped, birateral networks but also to networks containing distributed and/or non-birateral subnetworks. Digital computer program and example are discussed.

Further investigations of properties of series-parallel n-port networks and method of decomposition of whole network into subnetworks are remaining important problems.

The method presented here has been applied to almost all practical problems and is useful in every case.

VII ACKNOWLEDGMENT

The authors express their gratitude to Mr. Masanori Kogo, Manager of Components Engineering Department for his constant encouragements at all stages of this work. They also wish to thank T. Yano and the colleagues of the Network Design Group of the Company for their valuable suggestions.

REFERENCES

(1) T. Tsuchiya and H. Watanabe A Method of Network Analysis by Digital Computer Monograph of C.T.T.G. of IECE of Japan (in Japanese)

(2) T.R. Bashkow A Note on Ladder Network Analysis PGCT vol. 8 No.2 June 1961.

(3) Y. Oono Formal Realizability of Linear Networks, Proc. of the symposium on active networks and feedback systems, Polytechnic Institute of Brooklyn, 1960.

(4) R.J. Duffin Topology of Series-Parallel Networks Journal of Math. Analysis and Applications 10, 1965.

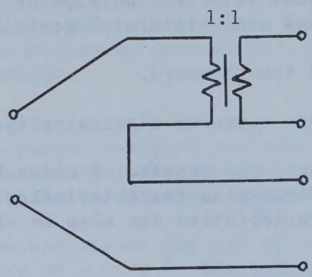


Fig.1 Series connection of two ports

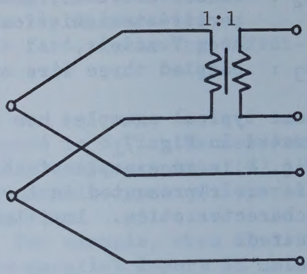


Fig.2 Parallel connection of two ports

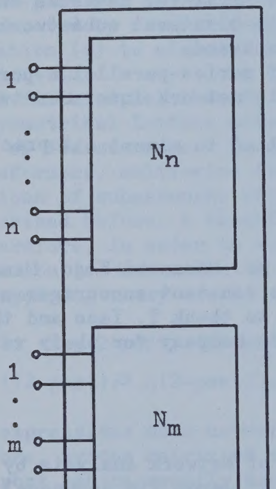


Fig.3 Connection of n-port and m-port makes (n+m)port

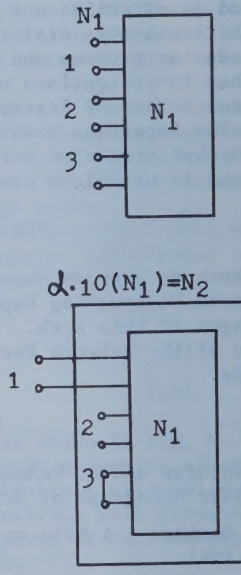

N_1
1
2
3
N_1
α·1O(N_1)=N_2
1
2
3
N_1

Fig.4 An example of unary operation of 3-port

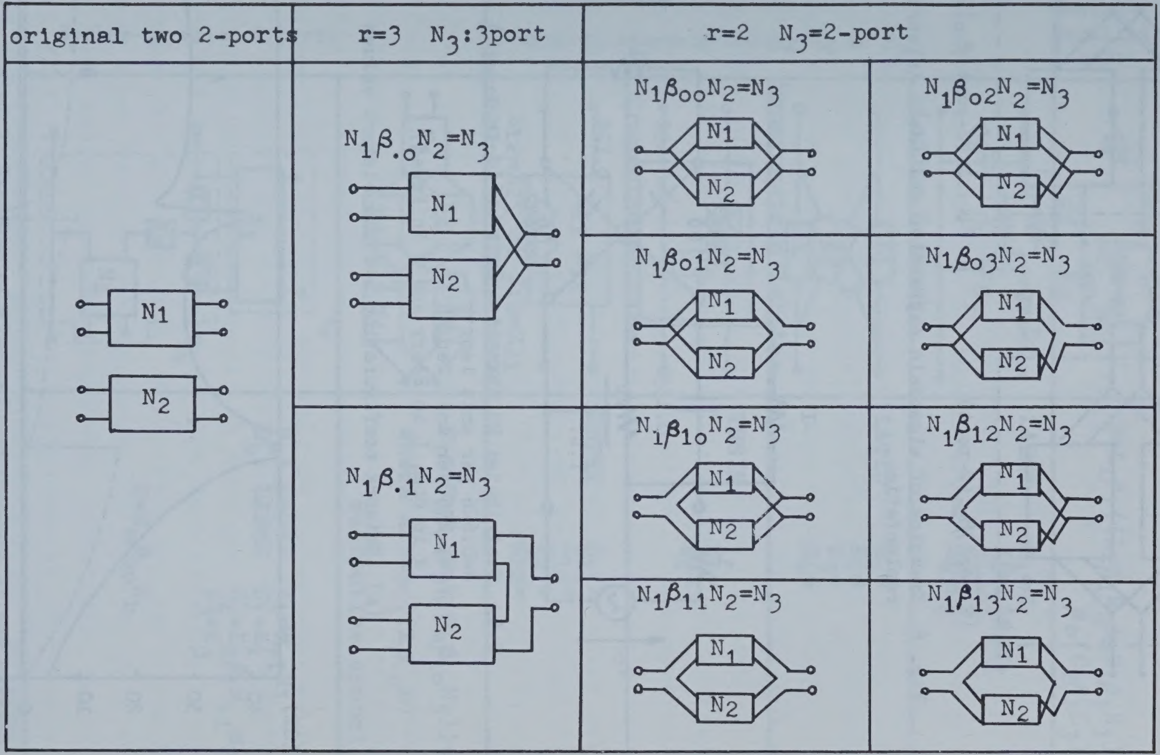
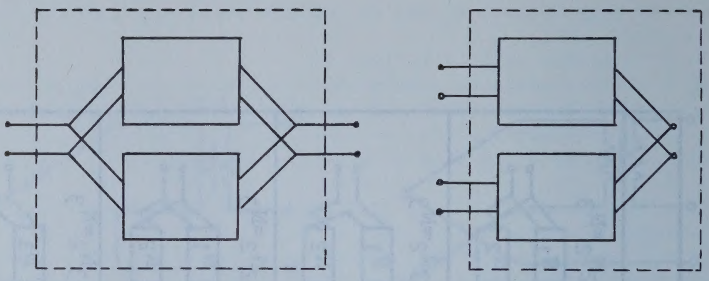
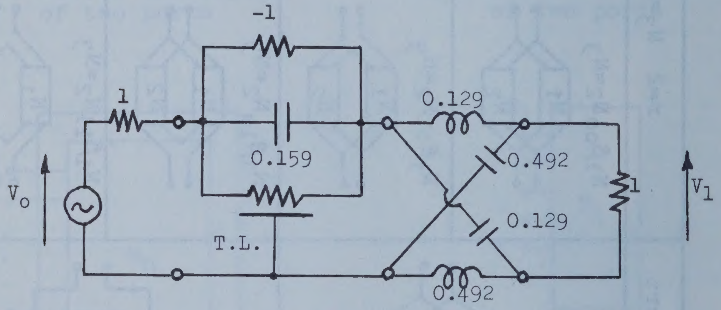

original two 2-ports
r=3 N$_3$:3port
r=2 N$_3$=2-port
N$_1$
N$_2$
$N_1\beta_{.o}N_2=N_3$
$N_1\beta_{.1}N_2=N_3$
$N_1\beta_{oo}N_2=N_3$
$N_1\beta_{o2}N_2=N_3$
$N_1\beta_{o1}N_2=N_3$
$N_1\beta_{o3}N_2=N_3$
$N_1\beta_{1o}N_2=N_3$
$N_1\beta_{12}N_2=N_3$
$N_1\beta_{11}N_2=N_3$
$N_1\beta_{13}N_2=N_3$

Fig.5 Some examples of binary operations with n=2



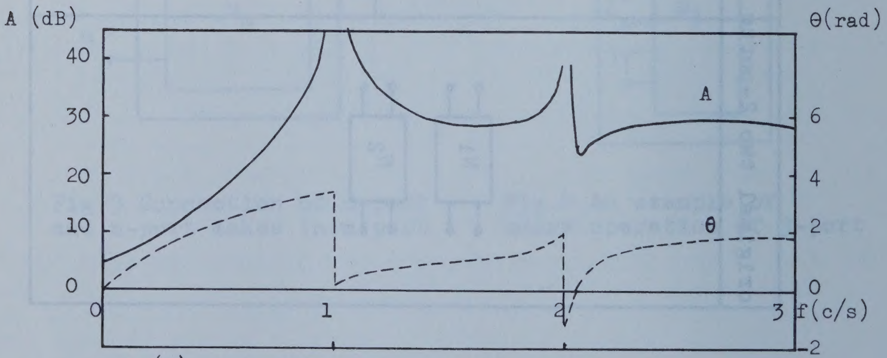
(2-port)β_{oo}(2-port) α..1 (2-port)β.o(2-Port)

Fig. 6 Examples of algebraic expression and their network replentation



T.L. Distributed RC transmission line (unit length)
R=0.626 per unit length
C=1.255 per unit length
(R in Ohm, L in Henry and C in Farad)

(a) Network configuration and the element values



(b) Attenuation and phase characteristics

Fig. 8 An example

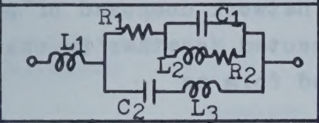
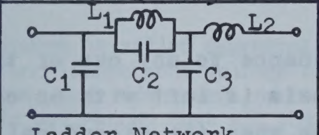
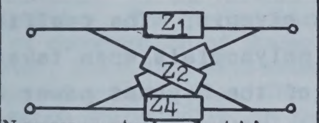
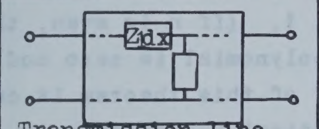
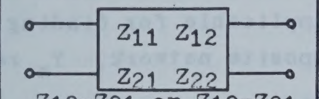
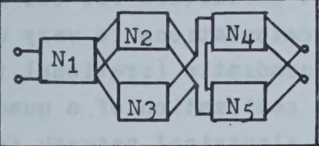
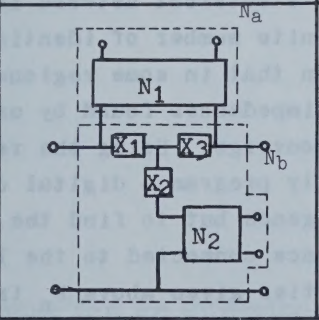
Type	Network Representation	Algebraic Expression (computer input data)
Series-parallel 1-port	 L1 R1 C1 L2 R2 C2 L3	$Z=L_1\beta_1((L_2\beta_1R_2\beta_oC_1\beta_1R_1)$ $\beta_o(C_2\beta_1L_3))$
Typical elements belonging to S_2	 L1 L2 C1 C2 C3 Ladder Network	LADEY $Z_1=C_1$ $Z_2=L_1\beta_oC_2$ $Z_3=C_3$ $Z_4=L_2$
	 Nonsymmetric lattice	LATIC $Z_1=$ $Z_2=$ $Z_3=$ $Z_4=$
	 Transmission line	TRSIN $Z_1=$ $Z_2=$
	 Z11 Z12 Z21 Z22 Z12=Z21 or Z12=Z21 non-bilateral network	ZMTRX $Z_{11}=$ $Z_{12}=$ $Z_{21}=$ $Z_{22}=$
	 N1 N2 N3 N4 N5	$N_1\pi(N_2\beta_{oo}N_3\pi(N_4\beta_{1o}N_5))=N$ where $A\pi B=\alpha_{.1}(A\beta_{.o}B)$ . . . subroutine symbol
Series-parallel 2 ports	 Na N1 X1 X3 Nb X2 N2	N_a: LADEZ $Z_1=\alpha_{.1}N_1$ N_b: LADEZ $Z_1=X_1$ $Z_2=X_2\beta_1\alpha_{.1}N_2$ $Z_3=X_3$ $N=N_a\beta_{oo}N_b$

Fig. 7 Examples of algebraic expressions versus network representations

NATURAL PROPERTIES OF CASCADED IDENTICAL SYMMETRIC NETWORKS

JOSEPH WOOD ROGERS
Electrical Engineering Department, Bucknell University,
Lewisburg, Pennsylvania, 17837.

The input impedance of a network composed of n identical 2-port symmetric networks connected together in cascade can be expressed in very simple closed form as

$$Z_n = z_{11}\ F_n(z_{11}\ y_{11})$$

in which z_{11} is the input impedance to any one of the networks when the second pair of terminals is left with an open circuit and y_{11} is the input admittance when the second pair of terminals is connected with a short circuit. The coefficients of the numerator and denominator polynomials when taken alternately starting with the coefficient of the highest power of $z_{11}y_{11}$ in the numerator is the sequence of binomial coefficients 1, n, n(n-1)/2, n(n-1)(n-2)/6 ... n, 1. (If n is even, the last coefficient of the denominator polynomial is zero and is excluded from the sequence.) The proof of this theorem is carried out using the principle of mathematical induction.

The results are equally applicable for finding the short circuit admittance for the composite network. Y_n replaces Z_n and y_{11} is interchanged with z_{11}.

Knowing the overal circuit parameters for the composite network without long involved calculations is very useful in the analysis and synthesis of quadratic irrational impedances. A sufficient condition for the realization of a quadratic irrational function as a passive electrical network is that the network be composed of an infinite number of identical symmetric networks. It is well known that in some regions of the complex plane the sequence of impedances found by using finite numbers of networks does not converge. Using the results obtained above one can efficiently program a digital computer not only to find regions of convergence but to find the effects of a different terminating impedance connected to the last network.

In addition to the properties given above F_n is a positive real function. This statement can be proved using R-C to L-C network conversion theorems. Also when n is even $F_n(1/x)$ is the same as $F_n(x)$; when n is odd $F_n(1/x)$ is the same as $1/F_n(x)$.

LINEAR DEGENERATE RC-NETWORK

SIK-KEE AU
IBM Corporation
San Jose, California

RUEY-WEN LIU*
University of California
Berkeley, California

ABSTRACT

A linear network is called degenerate if it has loops which consist only capacitors and independent voltage sources, and/or cut-sets which consist only inductors and independent current sources. If a network is not degenerate, it is well-known that all the capacitive voltages and inductive currents are continuous. If the network is degenerate, then we may encounter discontinuities. The discontinuous voltages and currents can be solved by means of conservation of charge and flux linkage or generalized functions. However, there is a well-known discrepency in energy loss which has not been answered analytically. The purpose of this paper is to explain this phenomenon analytically, in the spirit of lumped networks.

---

* On leave of absence from the University of Notre Dame, Notre Dame, Indiana September, 1965 - June, 1966.

DESIGN OF ONE-PORTS USING ONLY RESONATORS

JOHN F. O'NEILL
Laboratory for Electroscience Research
Department of Electrical Engineering
School of Engineering and Science
New York University, Bronx, New York 10453

I. PHYSICAL RESONATOR SYSTEMS AND SPECIAL DESIGN METHODS

It is well known that a stable bandpass resonance can be achieved in quartz and other piezolectric materials. A recent development makes use of the thickness twist mode of vibration in thin wafers to place several resonators on the same crystal.[1,2,3] Because the energy is largely trapped in the region between the exciting electrodes,[4] a single mode system with response free of harmonic and other spurious responses results. By variation of the spacing between pairs of electrodes the coupling among adjacent resonators can be adjusted to a desired value, or made neglibible. By this means it is possible to build, in the appropriate frequency range, useful networks which are compatible with integrated circuits in size, reliability, and cost. These devices are presently for application between about 1 MHz and 50 MHz.

The development of such resonators leads to the search for new network design methods using only resonant circuits in simple interconnections. Such a study should have interest independent of the physical system under discussion, since a network using even conventional element resonators can be pre-aligned before assembly can be compensated for temperature and aging one resonator at a time, and can be repaired easily, since the breakdown of one resonator element will usually give a localized error in the frequency response. Thus, an all-resonator type of network may be superior for practical application to an otherwise similar network with the same response.

At frequencies below those at which the crystal wafer technology can be used, various types of integrated circuit configurations using operational amplifiers can be used to synthesize resonant circuits. Low-pass all-resonator microstructures could be built in this way.

For application, transfer functions are usually needed to give sufficient frequency selectivity. In this paper is discussed driving point design only. One-ports deserve study for theoretical reasons principally, but they also have application as frequency sensitive amplifier interstages and loads. Admittances in proper form (low pass or bandpass) only are discussed; this causes essentially no loss of generality, and impedances proceed dually.

The resonator structure to be used for synthesizing admittances is a parallel connection of series resonators of the generalized 4-element type and special forms, which are shown as Fig. 1.

II. REALIZABILITY OF ADMITTANCE FUNCTION AS PARALLELED RESONATORS

A Test for Realizability as Paralleled Resonator Structure

We denote the complex frequency admittance of the general 4-element series resonator of Fig. 1 as $Y_r(s)$, where

$$Y_r(s) = \frac{1}{L} \frac{s+g/C}{s^2 + s(r/L+g/C) + \frac{1+rg}{LC}} \tag{1}$$

This function furnishes four degrees of freedom. The poles of $Y_r(s)$ appear as poles in the paralleled resonator structure, i.e., they are unchanged by the additional circuitry. The 2-element degenerate form furnishes two degrees of freedom and a single pole. To develop a rational admittance of order n, n even, n/2 4-element resonators furnish the required 2n degrees of freedom (n for the poles, n-1 for the n-1 zeros, and one for the coefficient). Similarly, an admittance of order m, m odd, requires $\frac{m-1}{2}$ 4-element resonators and one 2-element degenerate form. The important question remains whether the natural restrictions on the zeros, given the poles, for an all-resonator network permits the given Y(s) to be realized in this special form. Consider a pole pair at p and p* with partial fraction residues k and k*, where the asterisk denotes complex conjugation. We define Y_b by

$$Y_b(s) = \frac{k_1}{s-p} + \frac{k_1^*}{s-p^*} \tag{2}$$

Then the given Y(s) can be written

$$Y(s) = Y_a(s) + Y_b(s) \tag{3}$$

where $Y_n(s)$ has no pole at p. If the residues k and k* are to be suitable for realization as a resonator, we can equate Y_b and Y_r:

$$\frac{1}{L}\,\frac{s+g/C}{s^2 + s(r/L+g/C) + \frac{1+rg}{LC}} = 2 \text{ Re } k \,\frac{s - (\text{Re } p + \text{Im } p \text{ Im } k/\text{Re } k)}{(s-p)(s-p^*)} \tag{4}$$

Equation (4) must hold for all s, therefore Re k > 0. By examining the behavior of Y_r as r/L and g/C go to zero simultaneously, it can be seen that the product Im p$\cdot$Im k has the same sign as Re p (negative). For positive r, L, g, and C we know that $g/C \leq r/L + g/C = -2 \text{ Re } p$. Thus, conditions on the residue, k, can be written:

$$\text{Im } k \leq 0 \quad \text{Re } k > 0 \quad -\text{Re } p - \text{Im } k \text{ Im } p/\text{Re } k \leq -2 \text{ Re } p \tag{5}$$

or

$$\text{Im } k \leq 0 \quad \text{Re } k > 0 \quad \frac{-\text{Im } k}{\text{Re } k} \leq \frac{-\text{Re } p}{\text{Im } p} \tag{6}$$

This last is equivalent to

$$\text{Im } k \leq 0 \quad \text{Re } k > 0 \quad \tan(\arg k)\tan(\arg p) \leq 1 \tag{7}$$

One of each pair of conjugate poles in the given Y(s) must be tested for satisfaction of (7) before it can be stated that Y(s) is realizable as a structure of paralleled resonators. The test is essentially graphical, since the angle of the residue at any pole is simply measured by protractor from the pole-zero array of Y(s). Which pole to test first, i.e., which pole is most likely to fail, is not always apparent.

Application to Low Pass Maximally Flat Magnitude One-Port

To demonstrate the use of this test, let us consider a class of functions described by O'Neill and Ghausi.[5] These functions give maximally flat magnitude (MFM) and are easily shown to be realizable as driving point functions. The immittance function of complex frequency will be denoted by $W_n(s)$. Its magnitude in normalized form is given by

$$|W_n(\omega)|^2 = W_n(s)W_n(-s)\,\Big|_{s=j\omega} = \frac{1-\omega^{2n}}{1-\omega^{2n+2}} \tag{8}$$

Since a common factor cancels, $W_n(s)$ is of n^{th} order. Both poles and zeros lie on the unit circle, and the $j\omega$ axis pole is canceled by the $j\omega$ axis zero, as can be seen from (8). Fig. 2 shows a few of the low order pole-zero arrays for $W_n(s)$. The previous paper describes the frequency and time domain properties of the functions and mentions application as amplifier interstages. (As an interstage, the impedance $Z_n(s) \equiv W_n(s)$ can absorb both the shunt capacitance and shunt conductance at an interstage node, so it is suitable for use with triodes and transistors. Many other widebanding interstages do not have this capability.) The construction of Fig. 3 can be used to calculate the residue angle. For $W_n(s)$, but not as a general rule, the pole which most nearly fails the test (7) is that pole nearest the $j\omega$ axis. Thus, we define a reference pole, p_r, by $p_r \equiv p_n$. The angles to the m^{th} pole using this system of numbering is

$$\arg p_m = - \pi/2 - \frac{m\pi}{n+1} \qquad m = 1,2,\ldots,n \tag{9}$$

This can be seen from (8). Let us denote by β_n the angle to the reference pole, which will actually be used in the test. That is,

$$\beta_n = - \pi/2 - \frac{n\pi}{n+1} \tag{10}$$

The angle to the k^{th} zero is given by

$$\arg z_k = - \pi/2 - k\pi/n \qquad k = 1,2,\ldots,n-1 \tag{11}$$

The angle of a vector from the m^{th} pole to the reference pole is denoted αpm, and

$$\alpha p_m = - \pi/2 + \frac{\beta n}{2} + \frac{1}{2} \arg p_m = - \frac{3\pi}{4} + \frac{\beta n}{2} - \frac{m\pi}{2(n+1)} \tag{12}$$

The angle of a vector from the k^{th} zero to the reference pole is denoted by αz_k, and

$$\alpha z_k = - \frac{3\pi}{4} + \frac{\beta n}{2} - \frac{k\pi}{2n} \tag{13}$$

The angle of the residue at the reference pole, $\arg k_r$, is given by

$$\arg k_r \equiv \sum_{m=1}^{n-1} (\alpha z_m - \alpha p_m) = - \pi/4 \, \frac{n-1}{n+1} \qquad (14)$$

As mentioned, for the pole-zero array of $Y_n(s)$, all other poles have residue angle with smaller tangent than $\arg k_r$. Evaluating $\tan \beta_n \tan(\arg k_r)$, we find that the product is less than unity for $n \leq 4$, equal to unity for $n = 5$, and greater for $n > 5$, and the other conditions of (7) are satisfied. Direct examination of the special 3-element resonators of Fig. 1 shows that the condition $\tan(\arg p)\tan(\arg k) = 1$ corresponds to realization of the pole pair using one or the other of these special resonators. The fifth order MFM one-port is, thus, a slightly special case. The actual process of realization is nearly trivial if the factored function $Y(s)$ is given, since both the magnitude and angle of all residues can be got graphically. Fig. 4 gives a resonator realization of the third order MFM admittance.

It may be possible to realize some of the poles in a function using paralleled resonators and the remainder, if it is pr, using another method. The MFM one-port function obtained from (8) is one of few cases for which $\arg k_r$ can be simply evaluated algebraically.

Observations on Realizability

The previous discussion indicates that some given driving point functions of complex frequency are, indeed, realizable exactly as paralleled 4-element resonators. For practical application the 3-element rLC resonator of Fig. 1 may be the resonator which must be used, as would be the case for a crystal type of network. It is clear from the test (7) that essentially no arbitrarily given driving point function is realizable as paralleled rLC resonators. It is also apparent that nearly no bandpass function with high center frequency to bandwidth ratio is realizable as paralleled resonators of any type, since the term $\tan(\arg p)$ in the test (7) would be large for all poles, and the term $\tan(\arg k)$ cannot be expected to be near zero. Thus, it is clear that the paralleled resonator structure is very restrictive when attempting to realize a given pole-zero configuration. All is not gloom, however.

III. DESIGN OF SPECIFIED ADMITTANCE VS FREQUENCY CHARACTERISTICS USING PARALLELED RESONATORS

Although specific functions of complex frequency are generally not realizable using the practical rLC resonator, any specified magnitude vs frequency characteristic can be approached arbitrarily closely. To see this, reflect that any specified admittance shape has a unique real part which is positive or zero for all frequencies. For paralleled resonators, the real part of the structure is the sum of the real parts of all the resonator admittances. But the real part of the admittance of an rLC resonator can be shown, as, for instance, in Guillemin,[6] to approach an impulse function as the lossiness is reduced. Thus, a sufficiently large number of sufficiently lossless resonators can add arbitrarily closely to any specified single valued, real, function which is always positive or zero. The uniqueness of the relation between the real and imaginary parts of the driving point function guarantees that the proper magnitude, that of the desired admittance, results.

The effect of shunt parasitic capacitance, which would be present in a crystal network, places restrictions on the design, namely, the basic admittance level must be high enough to predominate over the capacitive susceptance in the frequency band of interest. The compensation of the design for shunt capacitance and an examination of out-of-band effects of this parasitic is carried out in some detail elsewhere.[7]

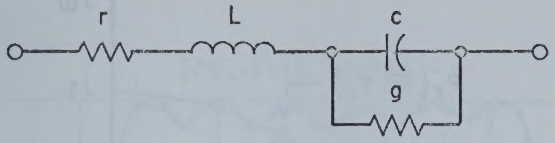
The problem of actually designing the resonators to match some specified shape can be attacked in several ways. Conceptually, it is most desirable to specify, say, only the number of resonators to be used, and leave all the other parameters (three per rLC resonator) to be varied in fitting the characteristic by some criterion. In practice, the available resonators may have restricted ranges of attainable Q, or some other restriction. Pre-specifying the poles of the network and varying only the zeros may be a useful attack on the design problem. If this scheme is used, only one degree of freedom per resonator, the coefficient 1/L, is available for variation. One criterion of match would be to equal the specified shape at as many different frequencies as there are resonators. (Both the specifying of the poles and the selecting of points to match are operations requiring judgment by the designer.) A simple algorithm for finding the coefficients in the above procedure can be found and implemented. Fig. 5 gives an example of an admittance found by such a procedure.

IV. ACKNOWLEDGMENT

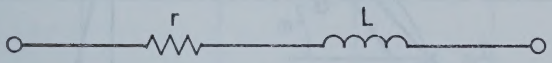
The author is supported by the doctoral program of Bell Telephone Laboratories. Part of the work described was supported by the Air Force Cambridge Research Laboratories under Contract AF19(628)-5537, to Prof. M. S. Ghausi.

REFERENCES

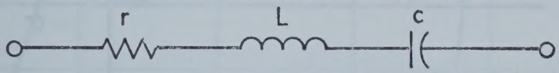
1. D. R. Curran, D. J. Koneval, "Miniature Ceramic Bandpass Filters," Proc National Electronics Conference, Vol 17, pp 514-520, 1961.
2. D. R. Curran, W. J. Gerber, "Low Frequency Ceramic Bandpass Filters," Proc National Electronics Conference, Vol 19, pp 108-113, 1963.
3. D. R. Curran, W. J. Gerber, D. J. Koneval, K. A. Pim, "Quartz Uniwafer Filters, A Thin Film Device," WESCON, Vol 2, 1964.
4. W. Shockley, D. R. Curran, D. J. Koneval, "Energy Trapping and Related Studies of Multiple Electrode Filter Crystals," Proc 17th Ann. Sym on Frequency Control, pp 88-126, May 1963.
5. J. F. O'Neill, M. S. Ghausi, "A Class of Realizable Maximally Flat Magnitude Driving Point Functions," NYU Technical Report 400-132, May 1966.
6. E. A. Guillemin, Synthesis of Passive Networks, New York: John Wiley and Sons, 1957.
7. J. F. O'Neill, M. S. Ghausi, "Design of Admittances Using Paralleled Resonators," NYU Technical Report 400-143, September 1966.



GENERAL 4-ELEMENT RESONATOR

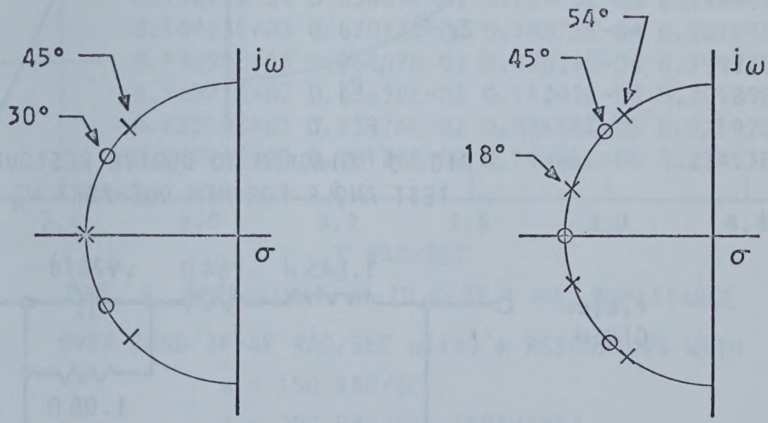


DEGENERATE FORM FOR LOW-PASS



3-ELEMENT PRACTICAL RESONATOR

FIG. 1 TYPES OF SERIES RESONATORS FOR ADMITTANCE DESIGN


45°
30°
jω
σ
54°
45°
18°
jω
σ

n = 3

n = 4

FIG. 2 TYPICAL POLE-ZERO ARRAYS FOR W_n (MTM ONE-PORT)

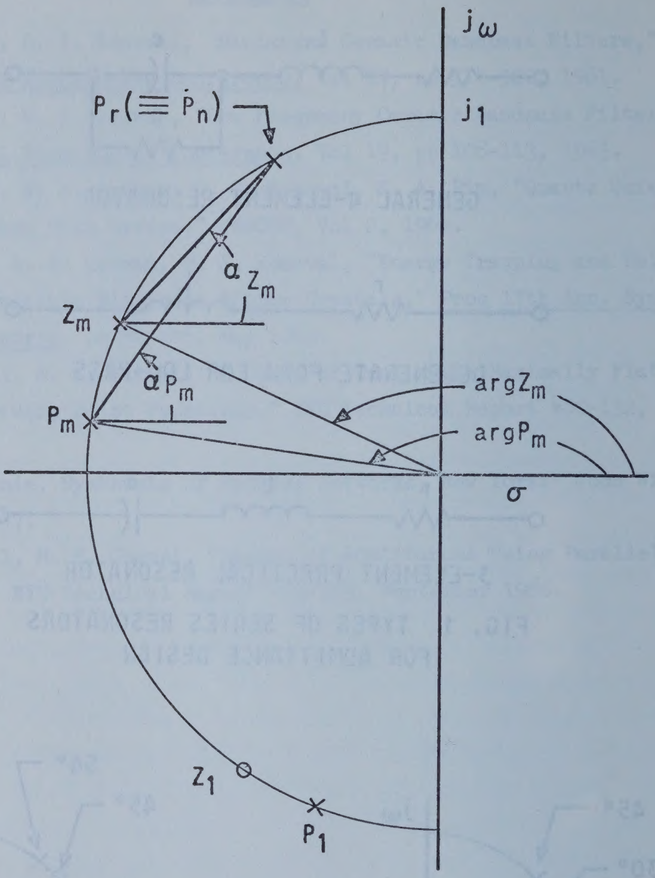

jω
j1
$P_r(\equiv \dot{P}_n)$
α_{Z_m}
Z_m
α_{P_m}
$\arg Z_m$
P_m
$\arg P_m$
σ
Z_1
P_1

FIG. 3 DIAGRAM TO DERIVE RESIDUE TEST ANGLE FOR MFM ONE-PORT W_n'

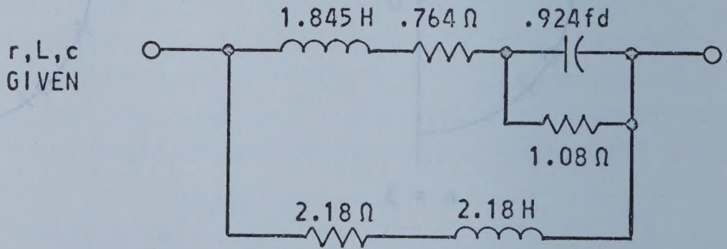
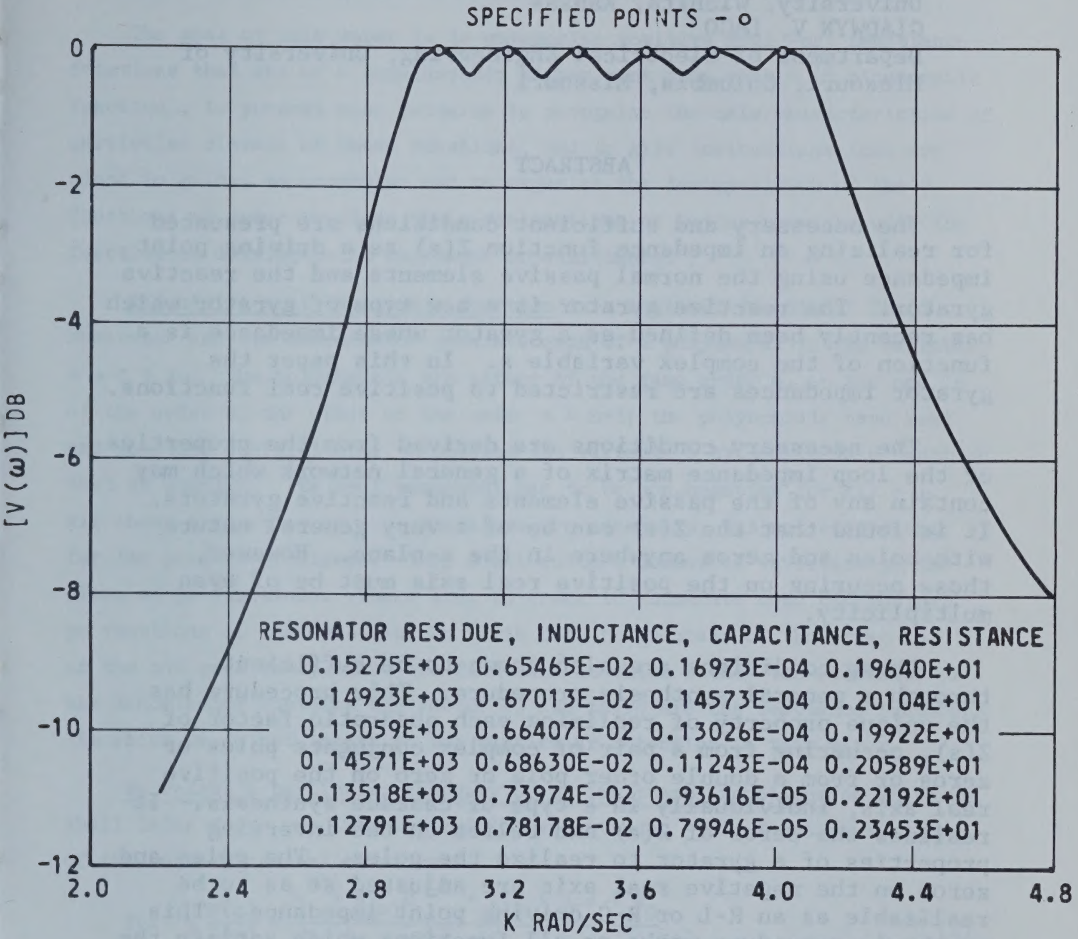

1.845 H
.764 Ω
.924 fd
r, L, c
GIVEN
1.08 Ω
2.18 Ω
2.18 H

FIG. 4 W_3 REALIZATION USING PARALLELED RESONATORS



RESONATOR RESIDUE,	INDUCTANCE,	CAPACITANCE,	RESISTANCE
0.15275E+03	0.65465E-02	0.16973E-04	0.19640E+01
0.14923E+03	0.67013E-02	0.14573E-04	0.20104E+01
0.15059E+03	0.66407E-02	0.13026E-04	0.19922E+01
0.14571E+03	0.68630E-02	0.11243E-04	0.20589E+01
0.13518E+03	0.73974E-02	0.93616E-05	0.22192E+01
0.12791E+03	0.78178E-02	0.79946E-05	0.23453E+01

FIG. 5 APPROXIMATION TO FLAT 1 MH$_o$ ADMITTANCE

OVER BAND 3K-4K RAD/SEC USING 6 RESONATORS WITH

a = 150 RAD/SEC

c = 200 RAD/SEC (SPACING)

SYNTHESIS OF DRIVING POINT IMPEDANCES WITH REACTIVE GYRATORS

ELMER A. HOYER*
Department of Electrical Engineering, Wichita State University, Wichita, Kansas
GLADWYN V. LAGO
Department of Electrical Engineering, University of Missouri, Columbia, Missouri

ABSTRACT

The necessary and sufficient conditions are presented for realizing an impedance function Z(s) as a driving point impedance using the normal passive elements and the reactive gyrator. The reactive gyrator is a new type of gyrator which has recently been defined as a gyrator whose impedance is a function of the complex variable s. In this paper the gyrator impedances are restricted to positive real functions.

The necessary conditions are derived from the properties of the loop impedance matrix of a general network which may contain any of the passive elements and reactive gyrators. It is found that the Z(s) can be of a very general nature with poles and zeros anywhere in the s-plane. However, those occuring on the positive real axis must be of even multiplicity.

These conditions are also shown to be sufficient through a general synthesis procedure. This procedure has the unique property of realizing each quadratic factor of Z(s), occurring from a pair of complex conjugate poles or zeros or from a double order pole or zero on the positive real axis, individually in a type of cascade synthesis. It realizes the zeros of Z(s) and relies on the inverting properties of a gyrator to realize the poles. The poles and zeros on the negative real axis are adjusted so as to be realizable as an R-L or R-C driving point impedance. This synthesis procedure works on all functions which satisfy the necessary conditions of which positive real functions are a special case.

---

*This paper represents a portion of the work in progress on a Ph.D. dissertation by the first author at the University of Missouri, Columbia, Missouri.

CLASSIFICATIONS AND PATTERNS OF RECOGNITION FOR POSITIVE REAL IMMITTANCE FUNCTIONS

KURT H. HAASE
Air Force Cambridge Research Laboratories, Bedford, Massachusetts

The goal of this paper is to categorize positive real (pr) immittance functions that are of a considerably higher rank than rank 4 (≡ biquadratic function), to present some patterns to recognize the main characteristics of particular classes of these functions, and to give instructions that are aimed to guide, to economize and to organize the decomposition of the functions in order to allow their realizations as ladder networks with the function as driving-point impedance or admittance.

A positive real immittance function is a rational function. Its numerator and its denominator consists each of a polynomial in the variable $s = \sigma + j\omega$. The polynomials are either of the same order n, or the one is of the order n, the other of the order $m = n+1$; the polynomials have real and positive coefficients. Replacing the variable s by $j\omega$ yields a function that is complex. Its real component has to be non-negative for all $\pm\omega$. All these properties of a pr functions are necessary, but not sufficient for its positive realness. They admit quite a number of varieties to the class of pr functions. There are, in order to enumerate some categories, pr functions of the even rank 2n (with both polynomials of the order n), of the odd rank 2n+1, there are functions in which either the numerator or the denominator contains the factor s, and there are the reactance functions -to which we do not intend to pay particular attention in this paper.

We select a particular category of pr immittance functions on which we shall later focus our interest and investigations. This function F(s) is of the rank 2n and has the form

$$F(s) = \frac{N(s)}{D(s)} = \frac{s^n + N_{n-1}s^{n-1} + \ldots + N_1 s + N_0}{s^n + D_{n-1}s^{n-1} + \ldots + D_1 s + D_0} \quad . \tag{1}$$

The function F(s) is an improper rational function and we shall further refer to it as a "nicely improper function" (nif-function) for the following reasons: none of its positive and real coefficients N_i and D_i in the sequence $i = 0, 1, \ldots, n$ is missing; particularly $N_n = D_n = 1$. It has no factor in common between its numerator and denominator. The lowest rank that F(s) can have is rank 0; in this event the function degenerates to the constant 1 because of the assumed normalization $N_n = D_n = 1$. Any pr function that is not a nif-function can be gradually reduced to a nif-function, as we will show later. This reduction - if contingent - is part of the realization procedure. Some further remarks on this procedure seem to be proper at this point.

To show the relations between the decomposition of an immittance function and its realization let f(s) be a pr function that may or may not be a nif-function at the moment. Assume that we are able to set

$$f(s) = f_1(s) + f_2(s) \tag{2}$$

where f_1 and f_2 are pr components, particularly f_1 a component of a very low rank, say rank 0, 1, or 2. Instead of decomposing f(s) we could also decompose

$$1/f(s) = 1/f_1^*(s) + 1/f_2^*(s) \tag{3}$$

with again $1/f_1^*$ a low-rank function. Following an earlier paper of the author$^{(1)}$ Eq. (3) can also be written in the script

$$f(s) = f_1^*(s) \oplus f_2^*(s) \quad . \tag{3a}$$

Having decomposed f(s), the next step is to decompose $f_2(s)$ and so forth until the entire function f(s) appears as a sum of low-rank components. When we alternate between decompositions according to (2) with decompositions according to (3a), + operator signs alternate with $\oplus$ operator signs in this sum. It has been shown in$^{(1)}$ that such a sum of low-rank immittance components can be identified with a ladder network of f(s) interpreted either as a driving-point impedance or admittance. As trivial examples consider Eqs. (2) and (3a). When we interpret f(s) as an impedance, the operator + in (2) means the series link between the impedances f_1 and f_2, and the operator $\oplus$ in (3a) means the parallel link between the impedances f_1^* and f_2^*. When we interpret f(s) as an admittance, the operator + means the parallel link between the admittances f_1 and f_2, and the operator $\oplus$ means the series link between the admittances f_1^* and f_2^*. The ladder structures in Figures 1a and 1b are somewhat more complicated. Let

$$f(s) = f_1 + f_2 \oplus f_3 + f_4 \oplus f_5 + f_6 \quad . \tag{4}$$

Interpreting all immittances in (4) as impedances yields the ladder in Fig. 1a and interpreting all immittances as admittances yields the ladder in Fig. 1b. Let now

$$f(s) = f_1 \oplus f_2 + f_3 \oplus f_4 + f_5 \oplus f_6 \quad . \tag{5}$$

Interpreting all immittances in (5) as admittances yields the ladder in Fig. 1a, and interpreting all immittances as impedances yields the ladder in Fig. 1b. Particularly these latter examples show that the design of the physical network can be completely separated from the decomposition of the immittance function. When we aim to a ladder structure, we only have to alternate between sequential decompositions of direct and of inverse immittance functions. The one or the other ladder structure is then only a question whether all components of the decomposition are interpreted as impedances or as admittances.

Some examples of decomposing pr, but not nif-functions are presented in TABLE 1. The function in row 1 has a pole, the function in row 2 has a zero for $s \to +\infty$. Both functions can be decomposed into a component of rank 1 and another component that is a nif-function F(s) multiplied with a positive constant factor. The function in row 3 has a pole, the function in row 4 has a zero for $s = 0$. Also these functions can be decomposed into a rank 1 function and a nif-function F(s). The original functions in rows 3 and 4 in the table are of the rank 2m; although similar functions that are of the ranks $2m \pm 1$ are not listed, the table gives sufficient information of how to decompose any non-nif-function according to the well known partial fraction decomposition so that finally a nif-function component is obtained. We shall now discuss the decompositions of nif-functions.

There are four categories of nif-functions. In a paper presented at the 1965 Allerton Conference the author(2) has discussed the realization of the biquadratic immittance function; this function is the lowest rank (rank 4) that a non-trivial nif-function can have. We have seen that there is one category for which the direct function F(s) and the inverse function 1/F(s) can be realized as an RC or as an RL ladder network. There is another category where the direct and the inverse function has to be expanded by the Brune procedure. There are a third and fourth category where either the direct or the inverse function has to be expanded according to Brune, but the other function has an RLC ladder realization. These categories can easily be recognized by investigating the sign polarities of a set of three test values. A very similar situation is contingent for a nif-function of any rank 2n, at least as far as the function is decomposed to another nif-function of the rank 2(n-2). A nif-function presented by (1) has the properties that

$$F(\infty) = 1, \qquad (6)$$

$$F(0) = N_0/D_0. \qquad (7)$$

We introduce the first test value T_0

$$T_0 = N_0 - D_0 , \qquad (8)$$

which merely by its polarity indicates the relation between the right sides of (6) and (7). Replacing s by $j\omega$ yields the complex function $F(j\omega)$ that has a real and an imaginary component according to

$$F(j\omega) = \text{Re } F(j\omega) + \text{Jm } F(j\omega) \ . \tag{9}$$

For convenience we introduce a new variable Ω as

$$\Omega = \omega^2 = - s^2 \ . \tag{10}$$

It is easy to show that

$$\text{Re } F(j\omega) = \frac{R(\Omega)}{P(\Omega)} = \frac{\Omega^n + R_{n-1}\Omega^{n-1} + \dots + R_1\Omega + R_0}{\Omega^n + P_{n-1}\Omega^{n-1} + \dots + P_1 s + P_0} \ , \tag{11}$$

is a rational function where normalized polynomials of the order n in Ω appear in the numerator and the denominator. The coefficients R_i and P_i however are not necessarily all positive. It can also be shown that

$$\text{Jm } F(j\omega) = - j\omega \frac{Q(\Omega)}{P(\Omega)} = -j\omega \frac{Q_{n-1}\Omega^{n-1} + Q_{n-2}\Omega^{n-2} + \dots + Q_0}{\Omega^n + P_{n-1}\Omega^{n-1} + \dots + P_0} \ , \tag{12}$$

where in $Q(\Omega)/P(\Omega)$ a non-normalized polynomial of the order n-1 in Ω is divided by the same polynomial $P(\Omega)$ as in (11). Also the coefficients Q_i are not necessarily positive. In (11)

$$R_i = N_iD_i + \sum_{k=1}^{j} (N_{i-k}D_{i+k} + N_{i+k}D_{i-k})(-1)^k \ , \tag{13}$$

in (11) and (12)

$$P_i = D_i^2 + 2 \sum_{k=1}^{j} D_{i-k}D_{i+k}(-1)^k \ , \tag{14}$$

and in (12)

$$Q_i = \sum_{k=1}^{j} (N_{i-1-k}D_{i+k} - N_{i+k}D_{i-1-k})(-1)^k \ . \tag{15}$$

The upper limit j in the summations (13), (14), and (15) is reached whenever a sub-index becomes either n or 0. There are simple rules of combination by which the coefficients R_i, P_i, and Q_i can be easily obtained from a tabulation of the coefficients N_i and D_i of F(s). These rules are independent on the magnitude of n. This is one of the advantages to focus the main interest on the first derived nif-function rather than to investigate an original function that is not a nif-function and therefore needs individual programming according to the afore mentioned varieties. Even if a function has to be tested on its positive realness, it is of advantage to reduce it first to the nif-function and to find out whether this function is pr. As far as positive realness is concerned, it is necessary that

$$\text{Re } F(j\omega) = \frac{R(\Omega)}{P(\Omega)} \geqq 0 \text{ for all } \quad 0 \leqq \Omega \leqq +\infty \quad . \tag{16}$$

By the real evaluations (6) and (7) also

$$\frac{R(\infty)}{P(\infty)} = 1 \ , \tag{17}$$

$$\frac{R(0)}{P(0)} = N_0/D_0 \ . \tag{18}$$

When we plot in a diagram Re $F(j\omega)$ vs $+\Omega$, the curve always starts at the point P_a and ends at the point P_1 when we procede along the $+\Omega$-axis. If T_0 defined in (8) is positive the point P_a with the abscissa $\Omega = 0$ is on a higher level than the point P_1 at the abscissa $\Omega = \infty$. The opposite is true when T_0 is negative. This is indicated in Figures 2a and 2b. The slope of the curve at P_a can be positive or negative. The direction of the slope can easily be recognized by the mere sign polarity of the second test value T_a that is defined as

$$T_a = D_0(N_0D_2 - N_2D_0) - D_1(N_0D_1 - N_1D_0) \quad . \tag{19}$$

The slope of $R(\Omega)/P(\Omega)$ vs Ω has the same polarity as the test value T_a, that is, the curve is ascending in direction of $+\Omega$ if T_a is positive. At the other end at P_1 where $\Omega \rightarrow +\infty$ the slope is horizontal. However, whether this point is approached from above or from below the ordinate 1 can be recognized by the mere sign polarity of a third test value T_1 that is defined as

$$T_1 = (N_{n-2} - D_{n-2}) - D_{n-1}(N_{n-1} - D_{n-1}) \ . \tag{20}$$

The curve approaches P_1 from below the ordinate 1 when T_1 is positive and from above when it is negative. Note that computing the test values does not need the coefficient R_i or P_i. According to Figures 2a,b the test values present valuable information on the characteristics of a nif-function. Let T_0 be positive. When $R(\Omega)/P(\Omega)$ vs Ω does not cross the ordinate 1, that is when the equation

$$\sum_{i=0}^{n-1} (R_i - P_i)\,\Omega^i = 0 \tag{21}$$

has no real and positive solution $\Omega = +\Omega_1$, then

$$\text{Re } F(j\omega) - 1 \geqq 0 \ , \tag{22}$$

and $F(s) - 1$ is still a positive real function; it is of the rank $2n-1$ and can be reduced to a nif-function of the rank $2n-2$. This however is certainly not possible when T_1 is positive. In this event it is certain that the curve crosses the ordinate 1 at least once and the direct function $F(s)$ has to be expanded according to Brune. Then we have to find the location Ω_m and the magnitude N_m of the absolute minimum of the curve Re $F(j\omega)$ vs Ω. Finding the minimum without pre-investigation as one of the solutions of the equation

$$\frac{d\,\frac{R(\Omega)}{P(\Omega)}}{d\,\Omega} = 0 \tag{23}$$

would be uneconomical. Equation (23) is identical with the equation

$$\sum_{i=0}^{2n-2} S_i\Omega^i = \sum_{i=0}^{n} P_i\Omega^i \sum_{i=1}^{n} iR_i\Omega^{i-1} - \sum_{i=0}^{n} R_i\Omega^i \sum_{i=1}^{n} iP_i\Omega^{i-1} = 0 \tag{24}$$

that is of the order 2n-2. Instead we better find in rough approximations the positive real solutions of Eq. (21) that is of the order n-1 only. These solutions are the crossings $+\infty$, Ω_{11}, Ω_{12}, ... of Re $F(j\omega)$ with the ordinate 1. The minimum then is located within $+\infty$ .... Ω_{11}, or Ω_{12}... Ω_{13} and so forth. When T_1 is negative and we want to make sure that the curve does not cross the ordinate 1 in the range $0 \leqq \Omega \leqq +\infty$ we only have the polynomial in (21) undergo the well known "STURM TEST".

A very similar situation we meet when T_0 is negative. Then the ordinate N_0/D_0 interchanges its role with the ordinate 1. Crossings of the ordinate N_0/D_0 are obtained by finding the positive and real solutions of the equation

$$\sum_{i=1}^{n} (D_0R_i - N_0P_i)\,\Omega^{i-1} = 0 \tag{25}$$

which as Eq. (21) is of the order n-1. The crossings when T_a negative are 0, Ω_{a1}, Ω_{a2}, ... and the minimum of Re $F(j\omega)$ is located within one of the ranges 0 ... Ω_{a1}, Ω_{a2} ... Ω_{a3} , and so forth. When T_a is positive and we do not expect any crossing, the polynomial in (25) has to undergo the STURM TEST for proof. The STURM TEST which can uniformly be programmed for nif-functions is also a partial test for positive realness when applied to the polynomial $R(\Omega)$.

REFERENCES

1. Haase, K. H., "A symbolic Notation Applied to Unbalanced Ladder Networks" NTZ, Vol 17, pp 572-578, 1964.
2. Haase, K. H., "A Classification of pr Biquadratic Immittance Functions and Their Realization Using a Perfectly Coupled and Center Branch Augmented Two-Port", Proc. 3rd Allerton Conf on Circuit and System Theory.

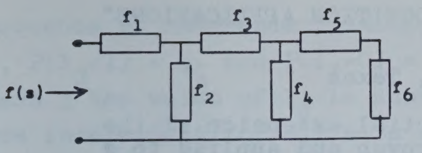

f(s)
f1
f2
f3
f4
f5
f6

FIGURE 1a

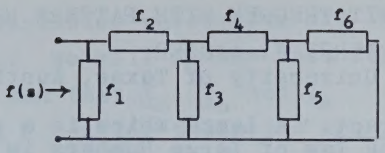

f(s)
f1
f2
f3
f4
f5
f6

FIGURE 1b

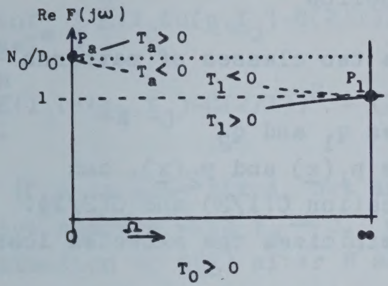


FIGURE 2a

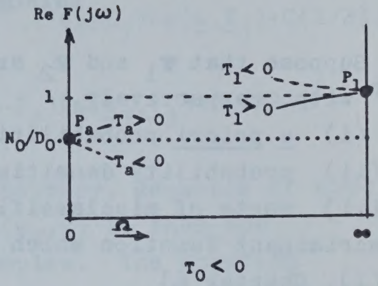


FIGURE 2b

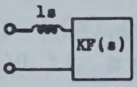
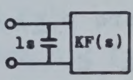
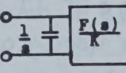
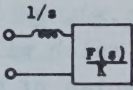
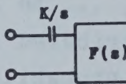
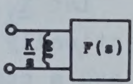
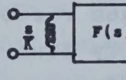
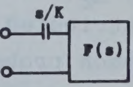
Decomposition of Immittance Functions	Interpretation as Impedance	Admittance
$\frac{n(s)}{D(s)} = s + KF(s)$ $K = n_n - D_{n-1}$ $N_n = 1,\ N_i = \frac{n_i - D_{i-1}}{K}\ (1 \leq i \leq n-1),\ N_0 = \frac{n_0}{K}$		
$\frac{N(s)}{d(s)} = \frac{1}{s} \oplus \frac{1}{K} F(s)$ $K = d_n - N_{n-1}$ $D_n = 1,\ D_i = \frac{d_i - N_{i-1}}{K}\ (1 \leq i \leq n-1),\ D_0 = \frac{d_0}{K}$		
$\frac{n(s)}{sD(s)} = \frac{K}{s} + F(s)$ $K = \frac{n_0}{D_0}$ $N_n = 1,\ N_i = n_{i+1} - KD_{i+1}\ (0 \leq i \leq n-1)$		
$\frac{sN(s)}{d(s)} = \frac{s}{K} \oplus F(s)$ $K = \frac{d_0}{N_0}$ $D_n = 1,\ D_i = d_{i+1} - KN_{i+1}\ (0 \leq i \leq n-1)$		
Definitions $m = n + 1$ $n(s) = s^m + n_{m-1}s^{m-1} + \ldots + n_1 s + n_0$ $d(s) = s^m + d_{m-1}s^{m-1} + \ldots + d_1 s + d_0$	$F(s) = \frac{N(s)}{D(s)}$ $N(s) = s^n + N_{n-1}s^{n-1} + \ldots + N_1 s + N_0$ $D(s) = s^n + D_{n-1}s^{n-1} + \ldots + D_1 s + D_0$	

TABLE I

A LIMIT THEOREM WITH PATTERN RECOGNITION APPLICATIONS*

TERRY J. WAGNER
University of Texas, Austin, Texas

Abstract. A Lemma which is a partial extension of the Strong Law of Large Numbers is proven and applied to a pattern recognition problem.

INTRODUCTION

Suppose that π_1 and π_2 are two classes of patterns in R^p with, respectively,

(i) *a priori* probabilities q_1 and q_2
(ii) probability densities $p_1(\underline{x})$ and $p_2(\underline{x})$, and
(iii) costs of misclassification $C(1/2)$ and $C(2/1)$.

A discriminant function which minimizes the expected loss is ([1], Chapter 6)

$$D(\underline{x}) = \begin{cases} \dfrac{C(2/1)q_1p_1(\underline{x}) - C(1/2)q_2p_2(\underline{x})}{p(\underline{x})} & \text{if } p(\underline{x}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $p(\underline{x}) = q_1p_1(\underline{x}) + q_2p_2(\underline{x})$ and we decide $\underline{x} \in \pi_1$ if $D(\underline{x}) \geq 0$ and $\underline{x} \in \pi_2$ if $D(\underline{x}) < 0$. We assume that q_1, q_2, $p_1(\underline{x})$ and $p_2(\underline{x})$ are unknown and that we have a function $u(\underline{a},\underline{x})$, $\underline{a} \in I^m$, $\underline{x} \in R^p$ (I is a closed, bounded interval of R) which will be used to approximate $D(\underline{x})$ by observing samples of π_1 and π_2 as follows.

Let $\underline{X}_1$, $\underline{X}_2$, ... be a sequence of independent, identically distributed random vectors with values in R^p and a common probability density p_1. Likewise let $\underline{Y}_1$, $\underline{Y}_2$, ... be a sequence of independent, identically distributed random vectors with values in R^p and a common probability density p_2. $\{\underline{X}_j\}$ then represents a sequence of independent samples from π_1 and $\{\underline{Y}_j\}$ a sequence of independent samples from π_2. For each j we indpendently observe $\underline{X}_j$ with probability q_1 or $\underline{Y}_j$ with probability q_2. If we put $I_j = 1$ if we observe $\underline{X}_j$ and $I_j = 0$ if we observe $\underline{Y}_j$ then I_1, I_2, ... is a

---

*Work supported by the National Science Foundation Grant GP-3012

sequence of independent random variables where, for all j, $P[I_j=1] = q_1$ and $P[I_j=0] = q_2$. We will assume that for each j the value of I_j is known and that $\underline{X}_m$, $\underline{Y}_n$ and I_ℓ are independent for all m, n and ℓ. After N samples are observed an $\underline{a}_N$ is chosen such that

$$\rho_N + \inf_{\underline{a}\in I^m} \frac{1}{N} \sum_1^N \{I_j[u(\underline{a},\underline{X}_j)-C(2/1)]^2 + (1-I_j)[u(\underline{a},\underline{Y}_j)+C(1/2)]^2\} \tag{1}$$
$$\geq \frac{1}{N} \sum_1^N \{I_j[u(\underline{a}_N,\underline{X}_j)-C(2/1)]^2 + (1-I_j)[u(\underline{a}_N,\underline{Y}_j)+C(1/2)]^2\}$$

where $\{\rho_N\}$ is some fixed, but arbitrary, sequence of non-negative numbers with $\rho_N \to 0$. $u(\underline{a}_N,\underline{x})$ is then the approximation of $D(\underline{x})$ after N samples. The question considered here is: how, and under what conditions, does $u(\underline{a}_N,\underline{x})$ approach $D(\underline{x})$ as N tends to infinity.

It should be stressed that what is being sought is a limit theorem. No algorithm for calculating $\underline{a}_N$ from the first N samples is given; in fact, what is desired is a result about what happens in the limit for any method of calculating $\underline{a}_N$ which satisfies (1). For specific algorithms, not all of which satisfy (1), see [2], [3], [4] and [5]. To the best of the writer's knowledge all algorithms for this problem always assume a $u(\underline{a},\underline{x})$ of the form $a_1\phi_1(\underline{x}) + \ldots + a_m\phi_m(\underline{x})$ where $\underline{a} = (a_1,\ldots,a_m)$ and $\phi_1,\ldots,\phi_m$ is a sequence of scalar functions.

THE MAIN RESULT

In what follows random vectors will be assumed to be defined on some fixed probability space $(\Omega,\mathcal{Q},P)$. The main result of this paper is the following

Theorem. Suppose that $u(\underline{a},\underline{x})$ is continuous on $I^m \times R^p$ where I is a closed, bounded interval of R. Then if $\int_{I_0^p} p(\underline{x})d\underline{x} = 1$ for some closed, bounded interval $I_0 \subset R$ and if $\{\underline{a}_N\}$ is any sequence of vectors from I^m such that (1) is satisfied for some $\omega \in \Omega$

$$\int_{R^p}[u(\underline{a}_N,\underline{x})-D(\underline{x})]^2 p(\underline{x})d\underline{x} \rightarrow \min_{\underline{a}\in I^m}\int_{R^p}[u(\underline{a},\underline{x})-D(\underline{x})]^2 p(\underline{x})d\underline{x}$$

with probability 1. If, in addition, there is a unique $\underline{a} = \underline{a}_o \in I^m$ minimizing $\int_{R^p}[u(\underline{a},\underline{x})-D(\underline{x})]^2 p(\underline{x})d\underline{x}$ then $\underline{a}_N \rightarrow \underline{a}_o$ with probability 1.

A statement of other results is given in the last section.

PROOF

Before proving the Theorem, four Lemmas are given with Lemma 4 being the key Lemma. Its proof depends only on the Strong Law of Large Numbers and routine analysis facts. The proofs of Lemmas 1, 2 and 3 are straightforward and are omitted. Below, $f(\underline{a},\underline{z})$ will always denote a nonnegative, continuous function defined on $I^m \times R^k$. $\{\rho_N\}$, as before, will denote an arbitrary, fixed sequence of nonnegative numbers with $\rho_N \rightarrow 0$.

Lemma 1. Let $\underline{Z}_1, \underline{Z}_2, \ldots$ be a sequence of random vectors with values in R^k. Then

$$W_N = \inf_{\underline{a}\in I^m} \frac{1}{N}\sum_1^N f(\underline{a},\underline{Z}_j) = \min_{\underline{a}\in I^m} \frac{1}{N}\sum_1^N f(\underline{a},\underline{Z}_j)$$

is a random variable.

Lemma 2. Let $\underline{Z}_1, \underline{Z}_2, \ldots$ be a sequence of independent identically distributed random vectors in R^k where $P[\underline{Z}_1 \in I_1^k] = 1$ for some closed, bounded interval $I_1 \subset R$. Let

$$\Omega_1 = \{\omega\in\Omega : \frac{1}{N}\sum_1^N f(\underline{a},\underline{Z}_j(\omega)) \rightarrow Ef(\underline{a},\underline{Z}_1) \text{ for all } \underline{a}\in I^m\}$$
$$\cap\{\underline{Z}_j \in I_1^k \text{ for all } j\}$$

Then Ω_1 is measurable (i.e., $\in \mathcal{Q}$) and has probability 1.

Lemma 3. Let $\underline{Z}_1, \underline{Z}_2, \ldots$ be a sequence of independent, identically distributed random vectors in R^k. If $Ef(\underline{a},\underline{Z}_1) < \infty$ for all $\underline{a}\in I^m$ then

$$\limsup_{N} W_N(\omega) \le \inf_{\underline{a} \in I^m} Ef(\underline{a},\underline{Z}_1) \text{ for all } \omega \in \Omega_1.$$

Lemma 4. Let $\underline{Z}_1$, $\underline{Z}_2$, ... be a sequence of random vectors satisfying the hypothesis of Lemma 2 and let

$$A = \{\underline{a} \in I^m : Ef(\underline{a},\underline{Z}_1) = \min_{\underline{a} \in I^m} Ef(\underline{a},\underline{Z}_1)\}$$

If $\{\underline{a}_N\}$ is any sequence of vectors from I^m for which

$$\rho_N + W_N(\omega) \ge \frac{1}{N} \sum_{1}^{N} f(\underline{a}_N, \underline{Z}_j(\omega)),\ N = 1,2,\ldots$$

is satisfied for some $\omega \in \Omega_1$ then

(a) $W_N(\omega) \to \min_{\underline{a} \in I^m} Ef(\underline{a},\underline{Z}_1)$ for all $\omega \in \Omega_1$,

(b) $d(\underline{a}_N, A) \to 0$ where $d(\underline{a}_N, A)$ is the distance between $\underline{a}_N$ and the set A,

(c) $Ef(\underline{a}_N,\underline{Z}_1) \to \min_{\underline{a} \in I^m} Ef(\underline{a},\underline{Z}_1)$, and

(d) $\frac{1}{N} \sum_{1}^{N} f(\underline{a}_N, \underline{Z}_j(\omega)) \to \min_{\underline{a} \in I^m} Ef(\underline{a},\underline{Z}_1)$ for all $\omega \in \Omega_1$.

proof of Lemma 4. We prove (a) first. From Lemma 2 and 3 it suffices to show that for any $\omega_o \in \Omega_1$

$$\liminf_{N} W_N(\omega_o) \ge \min_{\underline{a} \in I^m} Ef(\underline{a},\underline{Z}_1)$$

Let $\{\underline{a}_N\}$ be any sequence of vectors from I^m such that

$$\rho_N + W_N(\omega_o) \ge \frac{1}{N} \sum_{1}^{N} f(\underline{a}_N, \underline{Z}_j(\omega_o)) \text{ for all } N \tag{2}$$

and assume that N' is a subsequence such that $W_{N'}(\omega_o) \to \liminf_N W_N(\omega_o)$. Thus

$$\frac{1}{N'} \sum_{1}^{N'} f(\underline{a}_{N'}, \underline{Z}_j(\omega_o)) \to \liminf_{N} W_N(\omega_o)$$

Because $\{\underline{a}_{N'}\}$ is a bounded sequence in I^m it has a convergent subsequence $\{\underline{a}_{N''}\}$ with a limit in I^m, say $\underline{a}_{N''} \to \underline{a}_o \in I^m$. Since $\{\underline{a}_{N''}\}$ is a subsequence of $\{\underline{a}_N\}$

$$\frac{1}{N''} \sum_{1}^{N''} f(\underline{a}_{N''}, \underline{Z}_j(\omega_o)) \to \liminf_{N} W_N(\omega_o)$$

On the other hand

$$\left|\frac{1}{N''}\sum_{1}^{N''} f(\underline{a}_{N'''},\underline{Z}_j(\omega_o)) - \frac{1}{N''}\sum_{1}^{N''} f(\underline{a}_o,\underline{Z}_j(\omega_o))\right|$$

becomes arbitrarily small as $N'' \to \infty$ since $\omega_o \in \Omega_1$ and $f(\underline{a},\underline{z})$ is uniformily continuous on $I^m \times I_1^k$. Finally, since $\omega_o \in \Omega_1$

$$\frac{1}{N''}\sum_{1}^{N''} f(\underline{a},\underline{Z}_j(\omega_o)) \to Ef(\underline{a}_o,\underline{Z}_1)$$

so that $\liminf_N W_N(\omega_o) = Ef(\underline{a}_o,\underline{Z}_1)$ and (a) is proved.

The proof of (b) is established by contradiction. Suppose that $\{\underline{a}_N\}$ and ω_o satisfy (2). If $d(\underline{a}_N,A) \not\to 0$ there exists a $\delta > 0$ and a subsequence $\{\underline{a}_{N'}\}$ such that $d(\underline{a}_{N'},\underline{a}) \geq \delta$ for all N' and for all $\underline{a} \in A$ ($d(\underline{a}_{N'},\underline{a})$ is the distance between $\underline{a}_{N'}$ and $\underline{a}$). But from (a)

$$\frac{1}{N'}\sum_{1}^{N'} f(\underline{a}_{N'},\underline{Z}_j(\omega_o)) \to \min_{\underline{a}\in I^m} Ef(\underline{a},\underline{Z}_1)$$

By exactly the same argument used for (a) we can pick a convergent subsequence of $\{\underline{a}_{N'}\}$, say $\{\underline{a}_{N''}\}$, where $\underline{a}_{N''} \to \underline{a}_o$. But then $W_N(\omega_o) \to Ef(\underline{a}_o,\underline{Z}_1)$ and $\underline{a}_o \notin A$, a contradiction, and (b) is proven.

Finally, (c) follows from (b) and the fact that $f(\underline{a},\underline{z})$ is uniformly continuous on $I^m \times I_1^k$ and (d) follows by noting that for each $\omega \in \Omega_1$ every convergent subsequence of

$$\frac{1}{N}\sum_{1}^{N} f(\underline{a}_N,\underline{Z}_j(\omega))$$

has the same limit, namely, $\min_{\underline{a}\in I^m} Ef(\underline{a},\underline{Z}_1)$.

proof of the Theorem. After some straightforward manipulation it can be seen that

$$E\{I_1[u(\underline{a},\underline{X}_1)-C(2/1)]^2 + (1-I_1)[u(\underline{a},\underline{Y}_1)+C(1/2)]^2\} = \int_{R^p}[u(\underline{a},\underline{x})-D(\underline{x})]^2 p(\underline{x})d\underline{x} + c \qquad (3)$$

where c is a constant which does not depend on $\underline{a}$. Letting $\underline{Z}_j = (\underline{X}_j,\underline{Y}_j,I_j)$ and

$$f(\underline{a},\underline{z}) = i[u(\underline{a},\underline{x})-C(2/1)]^2 + (1-i)[u(\underline{a},\underline{y})+C(1/2)]^2 \qquad (4)$$

where $\underline{z} = (\underline{x},\underline{y},i)$, the proof follows from (3) and Lemma 4.

A FURTHER RESULT

By changing the definition of Ω_1 slightly and replacing the condition that the $\underline{Z}_j$ be bounded with probability 1 by:

$$
\begin{aligned}
&\text{(i)} \quad Ef(\underline{a},\underline{Z}_1) < \infty \text{ for all } \underline{a} \in I^m \\
&\text{(ii)} \quad |f(\underline{a},\underline{z})-f(\underline{b},\underline{z})| \leq h_1(\underline{a}-\underline{b})\phi_1(\underline{z}) + \ldots + h_r(\underline{a}-\underline{b})\phi_r(\underline{z}) \\
&\qquad \text{where, for all } j,\ h_j \text{ is continuous and } E|\phi_j(\underline{Z}_1)| < \infty
\end{aligned}
\tag{5}
$$

Lemmas 2,3 and 4 can be proven by essentially the same arguments as before. The Theorem then goes through as before if $u(\underline{a},\underline{x})$ yields an $f(\underline{a},\underline{z})$ (see(4)) satisfying (5). For example, if

$$u(\underline{a},\underline{x}) = \sum_1^r \varphi_j(\underline{a})\psi_j(\underline{x})$$

where, for all j, $\varphi_j(\underline{a})$ is a polynomial in the coordinates of $\underline{a}$ and $\int_{R^p}\psi_j^2 p(\underline{x})d\underline{x} < \infty$, then $f(\underline{a},\underline{z})$ satisfies (5).

REFERENCES

[1] T. W. Anderson, An Introduction to Multivariate Analysis, John Wiley and Sons, Inc., New York, 1958.

[2] M. A. Aizerman, E. M. Bravermann, L. Rozonoer, "The Method of Potential Functions for the Problem of Restoring the Characteristic of a Function Converter from Randomly Observed Points," Automation and Remote Control (English Translation) 25, No. 12, pp. 1546-1556, May, 1965.

[3] Y. C. Ho, C. Blaydon, "On the Abstraction Problem in Pattern Classification," Technical Report No. 476, Cruft Laboratory, Harvard University, Cambridge, Massachusetts, August, 1965.

[4] J. D. Patterson, B. F. Womack, "An Adaptive Pattern Classification System", Technical Report No. 9, Laboratories for Electronics and Related Science Research, University of Texas, Austin, Texas, August, 1965.

[5] R. L. Kashyap, C. C. Blaydon, "Recovery of Functions from Noisy Measurements Taken at Randomly Selected Points and Its Application to Pattern Classification", Proceedings of the IEEE, p. 1127, August 1966.

PERFORMANCE AND IMPLEMENTATION OF THE k-NEAREST NEIGHBOR DECISION RULE WITH INCORRECTLY IDENTIFIED TRAINING SAMPLES

A. W. WHITNEY and S. J. DWYER III
Department of Electrical Engineering
University of Missouri, Columbia, Missouri

ABSTRACT

A basic problem in pattern recognition is to develop procedures which classify observations such that a particular strategy is optimized. Bayes strategy is selected and a distribution-free decision rule, the k-nearest neighbor rule, is analized when each training sample is identified correctly with probability $\beta > 1/2$. In the two state equal loss problem, asymptotic bounds are given for the risk of the k-nearest neighbor rule, when $k = 1$ and $k \geq 1$ with k odd. Finally, a corollary is given which shows that the committee machine is in fact a nearest neighbor classifier under the appropriate metric.

THE k-NEAREST NEIGHBOR RULE

Consider the problem of developing decision rules which optimize Bayes strategy without requiring any assumptions concerning the parametric form of the underlying probability distributions. A particular distribution-free decision rule, the k-nearest neighbor (k-NN) rule, has been analized when the training samples are identified correctly[(2,3,4)].

Let a set of n state-measurement pairs $(\theta^1,x^1),\ldots,(\theta^n,x^n)$ be given as a set of n identified training samples. The x^i's are vector measurements which take values in a separable metric space upon which a metric ρ is defined. The θ^i's take the form of θ_1 or θ_2 when the measurement x^i is identified as being produced by θ_1 or θ_2, respectively.

When a new state-measurement pair is presented to the recognition system, only the measurement x is available to the decision rule and an estimate is desired of the state which produced that measurement. The k-NN rule uses the set of identified samples to classify the measurement x as follows:

Definition. The k-nearest neighbors to x are the set of k samples which are closest to x in the metric ρ. The k-NN rule assigns x to the state of the majority of the k-nearest neighbors[†].

Thus, the k-NN rule uses only the identification of the k-nearest neighbors and the identification of the remaining samples are ignored.

[†] Ties for the k-nearest neighbors may be resolved by randomization procedures.

Assumptions

The two allowable states of nature θ_1 and θ_2 are assumed to occur with a priori probabilities $Pr(\theta_1)$ and $Pr(\theta_2)$, respectively. Let $p(x|\theta_1)$ and $p(x|\theta_2)$ be conditional probability densities at x with respect to a σ-finite measure ν, such that $p(x|\theta_i)$ is the probability density of the vector measurement x given that the state θ_i, i=1,2, occurred. Ideally, each measurement x is classified by a Bayes decision rule which minimizes the expected loss. Thus, a loss, L_{ij} is associated with the decision that x was produced by θ_j when in fact, x was produced by θ_i. In the equal loss case, the losses are

$$L_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases}$$

when zero loss is associated with the correct decision or classification. Minimizing the expected loss is called Bayes strategy, which corresponds to minimizing the expected probability of error in the equal loss case. The minimum expected loss is also called the Bayes risk and the term risk will be employed.

The conditional Bayes risk, $r^*(x)$, in the two state equal loss case is

$$r^*(x) = \min [Pr(\theta_1|x), Pr(\theta_2|x)]. \tag{1}$$

Thus, the minimum risk, the Bayes risk, is given by $R^* = E[r^*(x)]$ where the expectation is with respect to the density $p(x)$.

Performance of the Single-NN Rule with Incorrectly Identified Samples

Usually, a sample identification system is used to identify the state of each of the n training samples. Let θ_i, i=1,2, be the actual input state and let $\hat{\theta}_j$, j=1,2, be the state which is provided by the sample identification system when the actual state is θ_i. It is assumed that the sample identification system has the property that

$$Pr(\hat{\theta}_i|\theta_i) = \beta > 1/2. \tag{2}$$

In this sense, each sample is said to be identified correctly with probability $\beta > 1/2$. Also, it is assumed that $p(x|\hat{\theta}_i, \theta_j) = p(x|\theta_j)$. Using these assumptions it follows that

$$p(\hat{\theta}_i, x) = p(\theta_i, x)\beta + p(\theta_j, x)(1 - \beta), \; i \neq j. \tag{3}$$

Note that $p(\hat{\theta}_i, x)$, i=1,2, are the apparent distributions which are used to train the k-NN rule but the actual distributions are $p(\theta_i, x)$, i=1,2. Even though these distributions differ, the following property is established for the two state equal loss case with $k = 1$ and $\beta > 1/2$.

Theorem 1. Let X be a separable metric space with a metric ρ. Let $p(x|\theta_1)$ and $p(x|\theta_2)$ be such that with probability one, x is either

i) a point of continuity of $p(x|\theta_1)$ and $p(x|\theta_2)$ or

ii) a point of nonzero probability measure.

Let each measurement be identified correctly with probability $\beta > 1/2$. Then the risk R of the single-NN rule has the asymptotic bounds $(n \to \infty)$

$$(1 - \beta) + (2\beta - 1)R^* \leq R \leq (1 - \beta) + (2\beta - 1)[2R^*(1 - R^*)]. \tag{4}$$

These bounds are as tight as possible.

Proof. Using conditional independence of the actual input state θ and the identified state of the nearest neighbor $\hat{\theta}$, the conditional single-NN risk $r(x,x_1^n)$ is given by

$$\begin{aligned} r(x,x_1^n) &= E[L(\theta,\hat{\theta}|x,x_1^n)] \\ &= Pr(\theta_1|x)Pr(\hat{\theta}_2|x_1^n) + Pr(\theta_2|x)Pr(\hat{\theta}_1|x_1^n) \end{aligned} \tag{5}$$

where the expectation is taken over θ and $\hat{\theta}$.

By the lemma of Cover and Hart(2), the single-nearest neighbor, x_1^n, to x converges to x with probability one; i.e. $x_1^n \to x$ with probability one.

It will now be shown that

$$Pr(\hat{\theta}_i|x_1^n) \longrightarrow Pr(\hat{\theta}_i|x) \tag{6}$$

with probability one. If x is a point of nonzero probability measure $\nu(x)$, then this point may be taken into account as follows. Let $\nu(x) > 0$; then

$$Pr(x \neq x_1^n) = [1 - \nu(x)]^n \longrightarrow 0, \; n \longrightarrow \infty. \tag{7}$$

But once $x_1^N = x$, $x_1^n = x$ for all $n \geq N$. Thus, $Pr(\theta_i|x_1^n) \to Pr(\theta_i|x)$ with probability one. For any remaining point x of zero probability measure, x is a continuity point of $p(x|\theta_1)$ and $p(x|\theta_2)$ with probability one. From Bayes theorem, $Pr(\theta_i|x)$ is continuous in $p(x|\theta_1)$ and $p(x|\theta_2)$. Hence x is a continuity point of $Pr(\theta_i|x)$ with probability one. But x_1^n, the single-nearest neighbor to x, converges to the random variable x with probability one. Therefore, $Pr(\hat{\theta}_i|x_1^n) \longrightarrow Pr(\hat{\theta}_i|x)$ with probability one, as was to be shown.

This convergence implies that with probability one

$$\begin{aligned} r(x,x_1^n) \longrightarrow r(x) &= Pr(\theta_1|x)[\beta Pr(\theta_2|x) + (1 - \beta)Pr(\theta_1|x)] \\ &\quad + Pr(\theta_2|x)[\beta Pr(\theta_1|x) + (1 - \beta)Pr(\theta_2|x)] \end{aligned}$$

$$\begin{aligned} r(x) &= 2\beta[Pr(\theta_1|x)Pr(\theta_2|x)] \\ &\quad + (1 - \beta)[(Pr(\theta_1|x))^2 + (Pr(\theta_2|x))^2]. \end{aligned} \tag{8}$$

The overall single-NN risk is by definition $R = \lim_{n\to\infty} E[r(x,x_1^n)]$ where the expectation is taken over x and x_1^n. But the conditional probability of error, and hence $r(x,x_1^n)$, is bounded; thus, by the dominated convergence theorem(5) $R = E[\lim_{n\to\infty} r(x,x_1^n)]$.

Applying this result to equation (8) implies

$$\begin{aligned} R &= E\{r(x)\} \\ &= E\{2\beta[\Pr(\theta_1|x)\Pr(\theta_2|x)] + (1-\beta)[\{\Pr(\theta_1|x)\}^2 + \{\Pr(\theta_2|x)\}^2]\} \\ &= E\{(1-\beta) + 2(2\beta-1)[\Pr(\theta_1|x)\Pr(\theta_2|x)]\} \\ &= E\{(1-\beta) + 2(2\beta-1)[r^*(x)][1-r^*(x)]\} \qquad (9) \\ &= (1-\beta) + 2(2\beta-1)[R^*(1-R^*) - \mathrm{Var}\,\{r^*(x)\}]. \end{aligned}$$

Hence, $R \leq (1-\beta) + (2\beta-1)[2R^*(1-R^*)]$ with equality iff $\mathrm{Var}\,\{r^*(x)\} = 0$, which holds iff $r^*(x) = R^*$ with probability one.

Now the lower bound is established by rewriting the equation (9) as follows

$$\begin{aligned} R &= E\{(1-\beta) + (2\beta-1)[r^*(x) + \{r^*(x)\}\{1-2r^*(x)\}]\} \\ &= (1-\beta) + (2\beta-1)R^* + (2\beta-1)E\{[r^*(x)][1-2r^*(x)]\} \\ &\geq (1-\beta) + (2\beta-1)R^* \qquad (10) \end{aligned}$$

with equality iff $E\,[r^*(x)][1-2r^*(x)] = 0$, which holds iff with probability one either $r^*(x) = 0$, or $r^*(x) = 1/2$.

It remains to be shown that these bounds are as tight as possible. Consider the case of $p(x|\theta_1) = p(x|\theta_2)$ to show that the upper bound is a least upper bound. For any a priori probabilities, $\Pr(\theta_1)$ and $\Pr(\theta_2)$,

$$R^* = \min\,\{\Pr(\theta_1), \Pr(\theta_2)\} \qquad (11)$$

and

$$\begin{aligned} R &= E\{r(x)\} \\ &= E\{\Pr(\theta_1|x)\Pr(\hat{\theta}_2|x) + \Pr(\theta_2|x)\Pr(\hat{\theta}_1|x)\} \\ &= E\{\Pr(\theta_1)[\beta\Pr(\theta_2) + (1-\beta)\Pr(\theta_1)] + \Pr(\theta_2)[\beta\Pr(\theta_1) + (1-\beta)\Pr(\theta_2)]\} \\ &= E\{(1-\beta) + (2\beta-1)2[\Pr(\theta_1)][1-\Pr(\theta_1)]\} \\ &= (1-\beta) + (2\beta-1)[2R^*(1-R^*)]. \qquad (12) \end{aligned}$$

Hence, for any R, $[0 \leq R \leq 1/2]$, a probability structure exists having Bayes risk R^* and single-NN risk $R = (1-\beta) + (2\beta-1)[2R^*(1-R^*)]$.

If the Bayes risk R^* is zero, the least upper bound is equal to the lower bound. Thus, the lower bound is a greatest lower bound when $R^* = 0$. Consider the case of $R^* > 0$ when both $\Pr(\theta_1)$ and $\Pr(\theta_2)$ are positive. This Bayes risk is obtained when

$$p(x|\theta_i) = \begin{cases} 1/\Pr(\theta_i),\ 0 < x < \Pr(\theta_i);\ i=1,2 \\ 0,\ \text{otherwise} \end{cases} .$$

Since both a priori probabilities are positive, the Bayes risk is

$$R^* = \min\,\{\Pr(\theta_1), \Pr(\theta_2)\}. \qquad (13)$$

Hence,

$$\begin{aligned} R &= E\{r(x)\} \\ &= E\{(1-\beta) + 2(2\beta-1)[\Pr(\theta_1|x)][1-\Pr(\theta_1|x)]\} \\ &= E\{(1-\beta) + 2(2\beta-1)[r^*(x)][1-r^*(x)]\} \\ &= E\{(1-\beta) + (2\beta-1)r^*(x) + (2\beta-1)[r^*(x)][1-2r^*(x)]\} \end{aligned}$$

$$R = (1-\beta) + (2\beta-1)R^*, \qquad (14)$$

since $r^*(x) = 1/2$ for $0 \leq x < \min\,[\Pr(\theta_1), \Pr(\theta_2)]$ and $r^*(x) = 0$ for $\min\,[\Pr(\theta_1), \Pr(\theta_2)] \leq x \leq \max\,[\Pr(\theta_1), \Pr(\theta_2)]$. Thus, the lower bound is a greatest lower bound because a probability structure exists with Bayes risk R^* and a single-NN risk $R = (1-\beta) + (2\beta-1)R^*$. A similar argument may be given if the a priori probabilities are stipulated. This completes the proof of the theorem.

Performance of the k-NN Rule with Incorrectly Identified Samples

Consider the k-NN rule for the two state, equal loss case with incorrectly identified samples, for a finite k. The argument given by Cover and Hart(2) may be modified to demonstrate that the k-th nearest neighbor to x converges to x with probability one, for fixed k. A similar argument may be applied to show that $\Pr(\hat{\theta}_i|x_1^n) \to \Pr(\hat{\theta}_i|x)$ with probability one.

In the two state, equal loss case with k odd, the k-NN risk $r(k,x)$ is

$$r(k,x) = \Pr(\theta_1|x) \sum_{i=0}^{(k-1)/2} \binom{k}{i}\{\Pr(\hat{\theta}_1|x)\}^i\{\Pr(\hat{\theta}_2|x)\}^{k-i}$$

$$+ \Pr(\theta_2|x) \sum_{j=(k+1)/2}^{k} \binom{k}{j}\{\Pr(\hat{\theta}_1|x)\}^j\{\Pr(\hat{\theta}_2|x)\}^{k-j}. \qquad (15)$$

Let $j = k-i$ in the second summation and applying the identity $\binom{k}{i} = \binom{k}{k-i}$ to the new index implies

$$r(k,x) = \sum_{i=0}^{(k-1)/2} \binom{k}{i}[\Pr(\theta_1|x)\{\Pr(\hat{\theta}_1|x)\}^i\{\Pr(\hat{\theta}_2|x)\}^{k-i}$$

$$+ \Pr(\theta_2|x)\{\Pr(\hat{\theta}_2|x)\}^i\{\Pr(\hat{\theta}_1|x)\}^{k-i}] \qquad (16)$$

where $$\Pr(\hat{\theta}_i|x) = \beta\Pr(\theta_i|x) + (1-\beta)[1-\Pr(\theta_i|x)]. \qquad (17)$$

It is desired to express $r(k,x)$ in terms of $r^*(x)$. This may be done by noting the symmetry of $r(k,x)$ in $\Pr(\theta_1|x)$ and $\Pr(\theta_2|x)$, which yields the same result for $r(k,x)$ whether $r^*(x) = \Pr(\theta_1|x)$ or $r^*(x) = \Pr(\theta_2|x)$. Thus suppressing the dependence on x

$$r(k,r^*) = \sum_{i=0}^{(k-1)/2} \binom{k}{i}\Big[r^*[\beta r^*+(1-\beta)(1-r^*)]^i[\beta(1-r^*)+(1-\beta)r^*]^{k-i}$$

$$+ (1-r^*)[\beta(1-r^*)+(1-\beta)r^*]^i[\beta r^*+(1-\beta)(1-r^*)]^{k-i}\Big]. \qquad (18)$$

Direct analysis of this expression is difficult, but the expected value of $r(k,r^*)$ can be bounded by the method employed by Cover and Hart(2). For this analysis, assume that k and β are fixed, such that $r(k,r^*)$ may be considered as a continuous function of r^* defined on the interval $[0,1/2]$ with

$$r(k,0) = \sum_{i=0}^{(k-1)/2} \binom{k}{i}(\beta)^{i}(1-\beta)^{k-i}. \tag{19}$$

For the given k odd and $\beta > 1/2$, define the convex upward function $C[k,r^*]$ by

$$\begin{aligned} C[k,r^*] &= \inf \phi(r^*) \\ &\qquad \phi(r^*) \geq r(k,r^*) \end{aligned} \tag{20}$$

for all convex upward functions $\phi(r^*)$. Also, define the convex downward function $G[k,r^*]$ by

$$\begin{aligned} G[k,r^*] &= \sup \psi(r^*) \\ &\qquad \psi(r^*) \leq r(k,r^*) \end{aligned} \tag{21}$$

for all convex downward functions $\psi(r^*)$. Thus, $G[k,r^*]$ is the largest convex downward function less than or equal to $r(k,r^*)$ and $C[k,r^*]$ is the smallest convex upward function greater than or equal to $r(k,r^*)$. By definition

$$G[k,r^*] \leq r(k,r^*) \leq C[k,r^*] \tag{22}$$

but these expressions are a function of r^* which is a function of x. Thus, taking the expectations over the induced random variable $r^*(x)$,

$$E\{G[k,r^*]\} \leq R(k) \equiv E[r(k,r^*)] \leq E\{C[k,r^*]\}. \tag{23}$$

However, $-G[k,r^*]$ and $C[k,r^*]$ are convex upward functions and applying Jensen's inequality$^{(5)}$

$$E\{-G[k,r^*]\} \leq -G[k,E(r^*)] = -G[k,R^*]$$

and

$$E\{C[k,r^*]\} \leq C[k,E(r^*)] = C[k,R^*]. \tag{24}$$

Therefore,

$$G[k,R^*] \leq R(k) \leq C[k,R^*] \tag{25}$$

which are the bounds of the k-NN risk since the dominated convergence theorem implies that $\lim_{n\to\infty} R_n(k) = R(k)$ where, $R_n(k)$, is the finite sample k-NN risk. These bounds can be shown to be the tighest possible by the same technique employed by Hart$^{(4)}$.

Theorem 2. Under the hypothesis of Theorem 1., the k-nearest neighbor risk is asymptotically bounded by

$$G[k,R^*] \leq R(k) \leq C[k,R^*] \tag{26}$$

where $G[k,R^*]$ and $C[k,R^*]$ are defined above. These bounds are as tight as possible.

The upper bound, $C[k,R^*]$, is plotted for $\beta = 1.0, 0.9, 0.8$ in Figure 1 and for several values of k. The lower bound, $G[k,R^*]$, is a straight line. This may be established by showing that $r(k,r^*) - y(k,r^*)$ is nonnegative, where $y(k,r^*)$ is a straight line joining the points $(0,r(k,0))$ and $(1/2,r(k,1/2))$, with $0 \leq r^* \leq 1/2$.

In addition, for any β with $1/2 < \beta \leq 1$, the difference $r(k,r^*) - y(k,r^*)$ converges monotonically to zero in k. Thus, $C[k,R^*]$ converges monotonically to $G[k,R^*]$ in k. Furthermore, $G[k,R^*]$ converges monotonically to R^* in k. Therefore, for any $\beta > 1/2$, $C[k,R^*]$ converges monotonically to R^* in k. This convergence of $C[k,R^*]$ to R^* in k, establishes the consistency of the k-NN rule when each sample is identified correctly with probability $\beta > 1/2$, $k \to \infty$, and $k/n \to 0$.

Similar properties may be established when β is a function of x. The upper bound is determined for that value of β_u such that

$$\beta_u = \inf_{x \in X} \beta(x) \tag{27}$$

and the lower bound is determined by that value of β_ℓ such that

$$\beta_\ell = \sup_{x \in X} \beta(x). \tag{28}$$

Also, the monotonic convergence in k of the upper bound, determined by β_u, to R^* may be established when $\beta_u > 1/2$.

IMPLEMENTATION OF THE k-NEAREST NEIGHBOR DECISION RULE

Two major problems in implementing the k-NN decision rule are first, the large number of samples and their identification which must be stored and second, the search through all these samples for each classification. In addition, the metric distance must be claculated between the measurement x to be classified and each of the n stored samples during the search procedure. Hence, the implementation depends on the metric chosen.

Several methods have been suggested which reduce the amount of storage required to retain some form of a minimum distance classifier. These methods usually consist of estimating a set of reference points and using these points as identified samples in a nearest reference point decision rule. The nearest mean and nearest mode decision rules are two examples of nearest reference point decision rules; Hart(4) has shown that these methods may be arbitrarily good or bad. Another rule has been suggested by Sebestyen(6) which he calls "adaptive sample set construction" but no analysis was given for its performance.

Implementation of the Nearest Neighbor Rule by the Committee Machine

The committee machine(1) is a two layer network of linear-threshold units consisting of an odd number of "committee" members which vote by majority. In the first layer, there are t (odd) linear-threshold units, each operating on the measurement vector x to produce a binary output. Thus, the committee machine is an analog to a committee where each linear-threshold unit is a committee member and produces a vote as it's output.

The second layer is a linear-threshold unit which operates on the component output of the first layer and produces a binary output. Hence, the second layer selects the majority vote of the first layer of committee members. Since the output is binary, a committee machine produces a dichotomy of the measurement space. An existence theorem was proven by Ablow and Kaylor(1) using a theoretical construction procedure.

Consider the performance of the committee machine. Since the committee machine partitions the measurement space, then the performance is a function of this partition. Each linear-threshold unit in the first layer may be thought of as dichotomizing the measurement space with a hyperplane. This hyperplane represents the set of points such that the argument of the signum function, which describes the linear-threshold unit, is zero. A linear-threshold unit associated with a hyperplane outputs a +1 (or -1) if the measurement point falls on the side of the hyperplane assigned to the state labeled +1 (or -1). Therefore, the committee machine partitions the measurement space into polyhedrons. The finite boundaries of these polyhedrons are formed by segments of the hyperplanes, where each hyperplane represents the decision boundary implemented by the corresponding linear-threshold unit. Each of these polyhedrons will be called a "cell". By construction, each cell contains only sample measurements which are identified as belonging to the same state. A cell is identified by noting the output of each of the t linear-threshold units of the first layer when any point in that cell is presented to the committee machine. If the first layer of t linear-threshold units are numbered, then each cell can be identified by a t sequence of +1's and -1's. The i-th element of the t sequence is determined by the output of the i-th linear-threshold unit, when a measurement point of that cell is presented to the linear-threshold unit.

When a new measurement point x is to be classified, the t sequence of the output of the first layer of linear-threshold units is noted. If the Hamming distance is zero between this t sequence and one of the t sequences associated with one of the cells, the measurement point actually lies in that cell. In fact, the Hamming distance between a cell containing a measurement point x and any cell Q is the minimum number of hyperplanes of the associated linear-threshold units that must be crossed in moving from the point x to the cell Q. Thus, the committee machine is a nearest neighbor classifier where the metric is the Hamming distance between a cell containing the measurement point and the constructed cells.

Corollary 1. The committee machine is a nearest neighbor classifier and has the same asymptotic bounds as the single-NN decision rule, when the Hamming metric is used to measure the distance between any cell containing a measurement point to be classified and the constructed cells.

Essentially, the committee machine acts as a single-nearest neighbor classifier. However, better bounds on the performance are obtained by using a k-nearest neighbor decision rule. Thus, it is desirable to have a committee machine also implement the k-nearest neighbor rule. The asymptotic performance of the k-nearest neighbor rule is obtained by using the committee machine to implement an approximation to the k-nearest neighbor rule.

Let the identified samples $x^1, x^2, \ldots, x^n$ be given. Let the identification provided for the samples represent the first label of identification. Construct a second label of identification using the k-nearest neighbor rule to classify the measurement x^i according to the first label with x^i removed from the sequence. Replace x^i and perform this operation for each sample measurement to form a complete set of second labels. Then the single-NN rule based on the second set of labels for the samples has the same asymptotic performance as the previous k-NN rule. Clearly the committee machine can implement the single-NN rule based on the second set of labels and hence has the same asymptotic performance as the k-NN decision rule.

If there are n identified training samples, then the committee machine requires at most n first layer linear-threshold units(1). The exact number of linear-threshold units required depends on the identified samples and the procedures employed to construct the committee machine. Unfortunately, no simple and effective construction procedures for the committee machine are known to the authors.

ACKNOWLEDGEMENTS

This paper represents a portion of the research work in progress on a Ph.D. dissertation by the first author at the University of Missouri, Columbia, Missouri. The first author wishes to acknowledge the support of a NASA Traineeship and the second author wishes to acknowledge the support of N.I.H. Grant MH 10936.

REFERENCES

(1) Ablow, C. M., Kaylor, D. J., "A Committee Solution to the Pattern Recognition Problem," IEEE Transactions on Information Theory, Vol. IT-11, No. 3, (July, 1965), pp. 453-455.

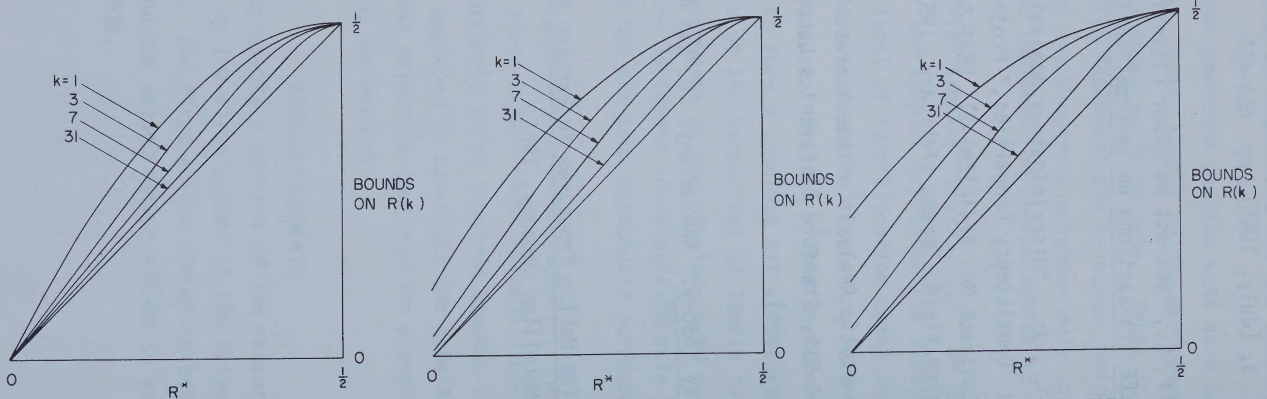
(2) Cover, T. M., Hart, P. E., "Nearest Neighbor Classification," to be published in IEEE Transactions on Information Theory.

(3) Fix, E., Hodges, J. L., Jr., "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," Project No. 21-49-004, Report No. 4, Contract No. AF41(128)-31, USAF School of Aviation Medicine, Randolph Field, Texas, (February, 1951).

(4) Hart, P. E., "An Asymptotic Analysis of the Nearest-Neighbor Decision Rule," SU-SEL-66-016, Stanford Electronics Laboratories, Stanford, California, (May, 1966).

(5) Loeve, M. Probability Theory. New Jersey: D. Van Nostrand Company, 1963.

(6) Sebestyen, G. Decision-Making Processes in Pattern Recognition. New York: The Macmillan Company, 1962.



PLOT OF ERROR BOUNDS FOR $\beta = 1.0$; $k = 1,3,7,31$

PLOT OF ERROR BOUNDS FOR $\beta = 0.9$; $k = 1,3,7,31$

PLOT OF ERROR BOUNDS FOR $\beta = 0.8$; $k = 1,3,7,31$

FIGURE 1

ON BAYESIAN LEARNING AND STOCHASTIC APPROXIMATION

Y. T. CHIEN and K. S. FU
Purdue University
Lafayette, Indiana

ABSTRACT

A general mathematical foundation for learning in an unknown stationary environment is established from the viewpoint of stochastic approximation theory. The existing learning techniques, based on the Bayesian type of inference, are first shown to fall into the general framework of stochastic approximation algorithms. As a consequence of this modeling, some convergence properties, optimal characteristics and possible improvements of the learning schemes are then derived with less effort and fewer restrictions. The results of this work therefore provide an alternative approach to the study of learning theory and suggest a new mathematical basis for the analysis and synthesis of learning systems in pattern recognition, automatic control and statistical communications.

I. INTRODUCTION

This paper introduces the basic algorithms in stochastic approximation theory and applies the viewpoints to unify the Bayesian and Bayesian-related learning procedures involved in many statistical learning system [1-6]. Particularly, it is intended to show that most of these learning procedures, which can be considered as successive estimation of parameters characterizing the underlying probability distributions, fall into the general structure of stochastic approximation initially devised by Robbins and Monro [8] and later extended, unified and generalized by Dvoretzky [9], among many other researchers. It is well known that stochastic approximation is concerned with the estimation algorithms which converge to the true value of some sought parameter when, due to the stochastic nature of the problem, the observations are inevitably noisy. The algorithms of most interest to our learning systems are those which possess the following properties: (1) they are self-correcting, that is, the error of estimates tends to vanish in the limit, (2) their convergence to the true value of the parameter is of some specific nature, for example, in mean square or with probability one. By illustrating that the supervised Bayesian learning procedures satisfy Dvoretzky's conditions, the convergence of successive estimates to their true parameters, in mean square and with probability one, is obtained by a simple application of Dvoretzky's theorem (see Appendix). As a result of this modeling, the optimal characteristic and possible improvement of the existing schemes can be derived with less effort and fewer restrictions. Finally, the effects of applying the viewpoints of stochastic approximation to nonsupervised learning are also discussed and a method of obtaining the parameter estimates through the moment estimators of the mixture distribution is presented.

II. NOTATIONS AND TERMINOLOGIES

A general rule is that capital letters are vectors and lower case letters are scalars. We define:

θ = Random variable (may be vector-valued) or parameters treated as random variable reflecting the unknown state of the probability distribution.

$X_1, X_2, \ldots, X_n$ = Sequence of independent identically distributed vector observations statistically related to θ (called learning observations or learning samples).

$p_o(\theta)$ = A priori probability density function for θ supposed to represent the initial knowledge about θ before any observation is made.

$p(\theta/X_1, X_2, \ldots, X_n)$ = A posteriori probability density function for θ, given the learning samples $X_1, X_2, \ldots, X_n$.

$p(X/\theta)$ = Probability density function of X, given θ.

III. SUPERVISED BAYESIAN LEARNING AS STOCHASTIC APPROXIMATION

The classical approach of supervised Bayesian learning is to assume that there exists an a priori density function $p_o(\theta)$ reflecting the initial knowledge about the unknown parameter θ and consider what happens to our knowledge of θ when the learning observations with known classification are made: $p_o(\theta)$ changes to the a posteriori probability density $p(\theta/X_1, X_2, \ldots, X_n)$ according to Bayes' theorem. For the learning procedures to be discussed [1,2], the iterative application of Bayes' theorem is accomplished by a fixed computational algorithm. This is made possible by carefully selecting a reproducing a priori density for the unknown parameter so that the a posteriori densities are members of the same family of the a priori density (i.e., the form of the density function is preserved). The learning processes are then reduced to the successive estimation of parameter values. We begin with the verifications that the estimation schemes thus developed are special algorithms of stochastic approximation. Optimal properties and possible improvements of these learning schemes are then derived as an immediate consequence of the verifications.

A. Learning the Parameters of Normal Distribution N(M,K) as Stochastic Approximation

Case (a). Unknown mean vector M and known covariance matrix K [1].

The parameter θ to be learned is M whose uncertainty can be reflected by assigning a proper reproducing a priori density $p_o(M)$, namely, Normal with mean vector M_o and covariance matrix $\emptyset_o$ where M_o represents the initial estimate of M and $\emptyset_o$ is the initial covariance matrix reflecting the uncertainty about M_o. By the reproducing property of normal density, it has been shown [1] that the a posteriori density $p(M/X_1, X_2, \ldots, X_n)$ is again Normal with M_o and $\emptyset_o$ replaced by their nth estimates M_n and $\emptyset_n$, respectively, where

$$M_n = (n^{-1}K)(\emptyset_o + n^{-1}K)^{-1} M_o + \emptyset_o(\emptyset_o + n^{-1}K)^{-1} < X > \tag{1}$$

$$\emptyset_n = (n^{-1}K)(\emptyset_o + n^{-1}K)^{-1} \emptyset_o \tag{2}$$

in which $< X > = \frac{1}{n} \sum_{i=1}^{n} X_i$. Equation (1) shows that the nth estimate of mean vector, M_n, can be interpreted as a weighted average of the a priori mean vector M_o and the sample information $< X >$. As $n \to \infty$, $M_n \to < X >$ and $\emptyset_n \to 0$ which means, on the average, the estimate M_n will approach M, the true mean vector of the normal distribution. Rewriting M_n in terms of M_{n-1} and X_n, we have

$$M_n = K(\emptyset_{n-1} + K)^{-1} M_{n-1} + \emptyset_{n-1}(\emptyset_{n-1} + K)^{-1} X_n \tag{3}$$

By adding and subtracting M_{n-1} and manipulating, equation (3) becomes

$$M_n = M_{n-1} + \emptyset_{n-1}(\emptyset_{n-1} + K)^{-1}(X_n - M_{n-1}) \tag{4}$$

which is a special algorithm of stochastic approximation. The nature of this algorithm can be easily seen in the special case when $\emptyset_o = \alpha^{-1}K$ and $\alpha > 0$ so that equation (4) simply reduces to

$$M_n = M_{n-1} + \frac{1}{n+\alpha}(X_n - M_{n-1}) = M_{n-1} + \gamma_n(X_n - M_{n-1}) \tag{5}$$

where $\gamma_n = 1/(n+\alpha)$ satisfying $\gamma_n > 0$, $\gamma_n \to 0$, $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$. To show that the successive estimate M_n is really a stochastic approximation algorithm satisfying Dvoretzky's conditions for convergence (see Appendix), we construct an underlying probability space in which X_n and M_n are random variables with $X_n = M + H_n$, where H_n is a k-vector $(\eta_n^1, \eta_n^2, \ldots, \eta_n^k)$ associated with the noise measurements at the nth estimation and satisfying $E\{\eta_n^i\} = o$ and $E\{(\eta_n^i)^2\} < \infty$, $i = 1,2,\ldots,k$, for all n. Then equation (5) becomes

$$M_n = (1-\gamma_n)M_{n-1} + \gamma_n M + \gamma_n H_n \qquad (6)$$

Let $T_n(M_1, M_2, \ldots, M_{n-1}) = (1-\gamma_n)M_{n-1} + \gamma_n M$ be the measurable transformation representing the "error-free" part of (6). We note

$$||T_n(M_1, M_2, \ldots, M_{n-1}) - M|| = ||(1-\gamma_n)M_{n-1} + \gamma_n M - M||$$
$$= (1-\gamma_n)||M_{n-1} - M|| = F_n||M_{n-1} - M|| \qquad (7)$$

where $F_n = 1-\gamma_n$ satisfies Dvoretzky's conditions (A1) and (A2) (see Appendix). Since $||M_o|| < \infty$ is assumed for the initial estimate and $E\{||H_n||^2\}$ is bounded above by B (i.e., $E\{||H_n||^2\} = \sum_{i=1}^{k} E\{(\eta_n^i)^2\} \leq B < \infty$), then $E\{||M_o||^2\} + \sum_{n=1}^{\infty} E\{||\gamma_n H_n||^2\} = E\{||M_o||^2\} + \sum_{n=1}^{\infty} \gamma_n^2 E\{||H_n||^2\} \leq E\{||M_o||^2\} + B\sum_{n=1}^{\infty} \gamma_n^2 < \infty$ which verifies condition (A4). Condition (A5) is satisfied for any measurable function $\varphi_n(M_1, M_2, \ldots, M_n)$. Thus, by Dvoretzky's theorem we have

$$\lim_{n \to \infty} E\{||M_n - M||^2\} = 0 \quad \text{and} \quad P\{\lim_{n \to \infty} M_n = M\} = 1 \qquad (8)$$

meaning that learning the mean vector of normal distribution with known covariance matrix is a special algorithm of stochastic approximation which, as a consequence of Dvoretzky's theorem, gives the convergence of the estimates to the true mean vector in mean square and with probability one.

Case (b). Unknown covariance matrix K and known mean vector (zero vector) [2].

In this case $\theta = K$ is the parameter to be learned. Let $K^{-1} = Q$ and assign the a priori density for Q to be the Wishart density with parameters (K_o, ν_o), where K_o is a positive definite matrix which reflects the initial knowledge of K and ν_o is a constant reflecting the confidence about the initial estimate K_o. It has been shown [2] that the a posteriori density of Q, $p(Q/X_1, X_2, \ldots, X_n)$, is again Wishart with parameters K_o and ν_o replaced by K_n and ν_n, respectively, as follows:

$$K_n = \frac{\nu_o K_o + n < XX^T >}{\nu_o + n} \qquad (9)$$

$$\nu_n = \nu_o + n \qquad (10)$$

in which $< XX^T > = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$. By adding and subtracting K_{n-1} as in case (a), one can easily show that equation (9) may be written as

$$K_n = K_{n-1} + \frac{1}{n+\nu_o}(X_n X_n^T - K_{n-1}) = K_{n-1} + \gamma_n(X_n X_n^T - K_{n-1}) \qquad (11)$$

where $\gamma_n = 1/(n+\nu_o)$ satisfies $\gamma_n > o$, $\gamma_n \to o$, $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$.

Note that equation (11) again depicts the standard algorithm of stochastic approximation in which the current estimate of the covariance matrix depends on the previous estimate and the sample correction term controlled successively by the numerical sequence γ_n. In order to verify that this learning algorithm also satisfies Dvoretzky's conditions for convergence, the elements of the covariance matrix can be treated individually and the same procedure as in case (a) should be carried out to show Dvoretzky's theorem applies for each and every element in the covariance matrix. As a result, the estimates obtained in (11) will converge to the true covariance matrix (i.e., for every element) in mean square and with probability one. The verification is very similar to that in case (a) and is omitted here.

Remark. It is also possible to show that Bayesian learning for both mean and covariance matrix of normal distribution [2] and for parameters of distributions other than Normal (e.g., binomial) form the stochastic approximation algorithms. The detailed analysis of there verifications may be found in [10].

B. Mean Square Error of Successive Estimates

It is often desirable in practice to determine the rate of convergence for a learning algorithm in terms of mean square error of the estimates. For the purpose of simplicity, consider the single-dimensional version of case (a) discussed above where learning the mean value m of a univariate normal distribution with known variance σ^2 may be considered as a special algorithm of stochastic approximation. Let m_n be the nth estimate of m. Then, according to (5), the learning algorithm is defined to be

$$m_n = m_{n-1} + \gamma_n(x_n - m_{n-1}) \tag{12}$$

where $x_n = m + \eta_n$ is the noisy observation at the nth stage, $E(\eta_n) = o$ and $E(\eta_n^2) = \sigma^2$. γ_n is in fact any numerical sequence satisfying the conditions previously described. To determine the mean square error for the estimates from (12), one can write

$$m_n - m = (1-\gamma_n)(m_{n-1}-m) + \gamma_n \eta_n \tag{13}$$

Squaring both sides and taking the conditional expection

$$E\{(m_n-m)^2/m_{n-1}\} = (1-\gamma_n)^2(m_{n-1}-m)^2 + \gamma_n^2 \sigma^2 \tag{14}$$

Taking the unconditional expectation and letting $V_n^2 = E\{(m_n-m)^2\}$, we have the following recursive relation for the mean square error:

$$V_n^2 = (1-\gamma_n)^2 V_{n-1}^2 + \gamma_n^2 \sigma^2 \tag{15}$$

C. Optimal Learning Sequence and Accelerated Algorithms

As a result of the above derivation, it is possible for us to obtain an optimal learning algorithm by choosing the γ_n sequence such that the mean square error at each and every stage is minimized. This optimization process can be achieved by setting the first derivative of V_n^2 with respect to γ_n equal to zero and solving for γ_n. The result is

$$\gamma_n = V_{n-1}^2/(\sigma^2 + V_{n-1}^2) \tag{16}$$

Let the initial mean square error be $V_o^2 = E\{(m_o-m)^2\}$. Using equations (15) and (16), we can iterate V_n^2 and γ_n alternately, obtaining the optimal sequence, γ_n^*, and the minimized mean square error, V_n^{2*}, as follows:

$$\gamma_n^* = \frac{1}{n+\sigma^2/V_o^2} = \frac{1}{n+\alpha}, \quad V_n^{2*} = \frac{\sigma^2}{n+\sigma^2/V_o^2} = \frac{\sigma^2}{n+\alpha} \tag{17}$$

where $\alpha = \sigma^2/V_o^2$ is a positive constant equal to the ratio of the variance of the noise to the variance of the initial estimate of the mean.

Note that the optimal sequence γ_n^* exhibited in (17) is exactly the same weighting sequence obtained in (5) of case (a) where the Bayesian learning of mean is found to be a special algorithm of stochastic approximation (cf. equations (4) and (5); also see [1], p. 61, where α is termed the dogmatism of the distribution). The minimized mean square error, V_n^{2*}, can also be shown to be the variance attained in the same learning procedure (cf. equation (2) which would be reduced to $\phi_n = \sigma^2/n+\alpha$ for the special case or see [1], p. 61). Thus it is reasonable to conclude that Bayesian learning algorithm is optimal in the sense of minimizing the mean square error. It should be pointed out, however, that this result is merely a consequence of optimizing the γ_n sequence for a specific stochastic algorithm, namely, algorithm using only the current sample x_n in the successive estimation (see equation (12)). Various algorithms in fact may be devised to produce even smaller error when statistics other than x_n are used. As an illustrative example, consider the statistic $< x > = \frac{1}{n} \sum_{i=1}^{n} x_i$ as our sample information in learning the mean value of a univariate normal distirubtion. Then a new algorithm can be defined as

$$m_n = m_{n-1} + \gamma_n(< x > - m_{n-1}) \tag{18}$$

Let λ_n denote the noise associated with $< x >$ so that $E(\lambda_n^2) = \frac{1}{n} E(\eta_n^2) = \sigma^2/n$ and let $< x >$ be replaced by $m + \lambda_n$. We can write

$$m_n = (1-\gamma_n)m_{n-1} + \gamma_n m + \gamma_n \lambda_n \tag{19}$$

Calculating the mean square error for the new algorithm one obtains

$$V_n^2 = (1-\gamma_n)^2 V_{n-1}^2 + \gamma_n^2 \sigma^2/n \tag{20}$$

Comparing equation (20) with equation (15), it is easily seen that the new algorithm converges more rapidly than the conventional Bayesian algorithm. The acceleration is obviously due to the reduced variance σ^2/n of the sample mean used in (20). To see quantitatively the efficiency of the two algorithms, mean square errors for both cases are computed based on their optimal γ_n sequence. The expressions γ_n^* and V_n^{2*} are given below.

Bayesian Algorithm (unaccelerated)

$$\gamma_n^* = \frac{1}{n+\alpha}, \quad V_n^{2*} = \frac{\sigma^2}{n+\alpha} \tag{21}$$

New Algorithm (Accelerated)

$$\gamma_n^* = \frac{n}{\frac{n(n+1)}{2} + \alpha}, \quad V_n^{2*} = \frac{\sigma^2}{\frac{n(n+1)}{2} + \alpha} \tag{22}$$

Expressions in (22) are obtained by minimizing V_n^2 of (20) with respect to γ_n. The successive mean square errors for both algorithms (assume $\sigma^2=1$, $\alpha=1$) are given in Table I. It is shown that at the 10th estimation, the error for the accelerated algorithm has already decreased to 1/56 while the error for the unaccelerated algorithm is still about 5 times bigger.

TABLE I COMPARISON OF MEAN SQUARE ERRORS ($\sigma^2=1$, $\alpha=1$)

No. of Stage n	0	1	2	3	4	5	6	7	8	9	10
Accelerated	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{7}$	$\frac{1}{11}$	$\frac{1}{16}$	$\frac{1}{22}$	$\frac{1}{29}$	$\frac{1}{37}$	$\frac{1}{46}$	$\frac{1}{56}$
Unaccelerated	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	$\frac{1}{9}$	$\frac{1}{10}$	$\frac{1}{11}$

IV. NONSUPERVISED LEARNING AND STOCHASTIC APPROXIMATION

Nonsupervised learning considers the problem where some fixed but unknown parameters in a mixture distribution are estimated from the learning samples with unknown classification. The solutions to this problem proposed in [3,4,6] are basically one of using the idea of Bayesian inference to compute the a posteriori densities for the unknown parameters. Unfortunately, the computation of these densities does not assume fixed algorithms and the property of convergence to the true parameters can not be achieved without making additional assumptions on the learning processes. In fact in many cases the solution, if exists, may not be unique. A significant aspect of the mixture approach to the nonsupervised learning, however, is that sufficient amount of a priori knowledge necessary for a unique solution to exist can be properly determined. It has been shown [6,7] that the class of mixture distributions which may have a unique solution is limited and whether they admit a unique solution will depend upon the identifiability[1] of the mixture distribution. We shall not be particularly concerned with the details of these existing techniques[2]. Instead, a unified method of finding the sought parameters in mixture distribution using the algorithm of stochastic approximation is presented. Examples of learning the mixture of univariate normal distributions are given to show under what conditions the solution may or may not be unique.

A. Formulation of the Nonsupervised Learning Problem

Let the learning process have the following set-up:

(1) Let $\Omega = \{\omega_i : i=1,2,\ldots,\ell\}$ be the set of ℓ classes of probability distributions and $P = \{P(\omega_i) = P_i : i=1,2,\ldots,\ell\}$ be the set of a priori probabilities, $\sum_{i=1}^{\ell} P_i = 1$.

(2) Each distribution is characterized by a probability distribution function, $P(X/\omega_i, S_i)$, or probability density function, $p(X/\omega_i, S_i)$, where S_i is the unknown parameter associated with ω_i.

(3) Unclassified learning samples are assumed drawn from the mixture distribution, i.e., $P(X/S,P) = \sum_{i=1}^{\ell} P_i \; P(X/\omega_i, S_i)$ where $P(X/S,P)$ denotes the mixture distribution function characterized by S and P.

(4) There exist unbiased estimates of certain statistics, $U = \{U(X)\}$, for the mixture (e.g., first moment, second moment, etc.). The functional relations between U and the parameters S and P, $F_i(U,S,P) = 0$, are known at each stage of the learning process.

(5) Additional equations relating the parameter sets, say $G_j(S,P) = 0$, are available to give unique solution for the unknown parameter S and P.

If (1) - (5) are satisfied with probability one, then the true parameters S and P are defined in the limit by the estimation equations $F_i = 0$ and $G_j = 0$. Moreover, the learning procedure is reduced to finding the unique solution for S and P through the estimation equations which are either given a priori or can be obtained from the successive estimates U using the stochastic approximation algorithms.

B. Examples of Stochastic Approximation Method

Problem statement: Let $\Omega = (\omega_1, \omega_2)$ be the set of two univariate normal distributions characterized by their means and variances, i.e.,

$$p(x/\omega_i, m_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{1}{2\sigma_i^2}(x-m_i)^2}, \quad i=1,2.$$

The mixture density is then given by

$$p(x/S,P) = P\,p(x/\omega_1, m_1, \sigma_1^2) + (1-P)\,p(x/\omega_2, m_2, \sigma_2^2) \tag{23}$$

in which $P = P(\omega_1)$ and $1-P = P(\omega_2)$ are the a priori probabilities. We wish to learn the unknown parameters $S = (m_1, m_2, \sigma_1^2, \sigma_2^2)$ and P through the unclassified learning samples $x_1, x_2, \ldots, x_n$, generated in some fashion from $p(x/S,P)$.

---

[1] The question of unique characterization for mixture distribution is termed identifiability in the literature (see, for instance, [7]).

[2] A discussion of the relation between the nonsupervised learning techniques and the stochastic approximation method may be found in [10].

Problem solution: Let the first, second and third moment of x with respect to $p(x/S,P)$ be computed from (23) as follows:

$$E(x) = P\, m_1 + (1-P)\, m_2 \tag{24}$$

$$E(x^2) = P(m_1^2 + \sigma_1^2) + (1-P)\,(m_2^2 + \sigma_2^2) \tag{25}$$

$$E(x^3) = P(m_1^3 + 3\, m_1\, \sigma_1^2) + (1-P)\,(m_2^3 + 3\, m_2\, \sigma_2^2) \tag{26}$$

Example (1). Consider P and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ are fixed and known. Let $m_2 = o$ and m_1 be the parameter to be learned. From (24), m_1 can be solved

$$m_1 = E(x)/P \tag{27}$$

It remains only to obtain the successive estimates of E(x). Stochastic approximation algorithm can be used to given asymptotically unbiased estimate of E(x), which will in turn given the unbiased estimate of m_1. Let $E_n(x)$ be the nth estimate of E(x) and define

$$E_n(x) = E_{n-1}(x) + \gamma_n\,(< x > - E_{n-1}(x)) \tag{28}$$

where γ_n is a positive numerical sequence satisfying the conditions previously described. Then by Dvoretzky's theorem we have

$$\lim_{n \to \infty} E\left\{(E_n(x) - E(x))^2\right\} = o \quad \text{and} \quad P\left\{\lim_{n \to \infty} E_n(x) = E(x)\right\} = 1$$

which imply

$$\lim_{n \to \infty} E\left\{\left(\frac{E_n(x)}{P} - m_1\right)^2\right\} = o \quad \text{and} \quad P\left\{\lim_{n \to \infty} \frac{E_n(x)}{P} = m_1\right\} = 1$$

That is, the true value of m_1 is learned in mean square and with probability one.

Example (2). Let $m_2 = o$ and let m_1, P and $\sigma^2 = \sigma_1^2 = \sigma_2^2$ be the parameters to be learned through the first three moments. Equation (24), (25) and (26) become

$$E(x) = P\, m_1 \tag{29}$$

$$E(x^2) = P\, m_1^2 + \sigma^2 \tag{30}$$

$$E(x^3) = P\,(m_1^3 + 3\, m_1\, \sigma^2) \tag{31}$$

Solving (29), (30) and (31) simultaneously gives

$$m_1^2 - 3\, E(x)\, m_1 + 3\, E(x^2) - E(x^3)/E(x) = o \tag{32}$$

Since (32) is a 2nd order equation in m_1, it can be readily shown that the discriminant is

$$9\, E^2(x) - 12\, E(x^2) + 4\, E(x^3)/E(x) \tag{33}$$

Substituting the relations (29), (30) and (31) for the moments, (33) becomes

$$m_1^2\, (3P-2)^2 \tag{34}$$

Since $m_1 \neq o$, we have the condition for

Unique solution: $P = 2/3$ (35)

Nonunique solution: $P \neq 2/3$ (36)

Thus, if (35) is satisfied for the mixture distribution then the parameters P, σ^2 and m_1 can be learned uniquely through (29), (30) and (31) by defining stochastic approximation algorithms to give asymptotically unbiased estimates of E(x), $E(x^2)$ and $E(x^3)$. Otherwise, the problem will have multiple solutions as shown in (36). In this case, the system does not learn anything either since it simply cannot distinguish one solution from the other and therefore is not assured that the solution chosen is the true one characterizing the mixture distribution.

V. CONCLUSIONS

We have shown in this paper that the learning techniques based on Bayesian inference can be interpreted within the general framework of stochastic approximation theory. The successive estimates of the unknown parameters are therefore proved to converge to their true values, in mean square and with probability one, by merely applying the results of Dvoretzky's theorem. Also indicated in this study is the rate of convergence of the Bayesian learning procedures in terms of mean square error. It is found that the evolution of stochastic approximation in Bayesian learning has permitted us to devise, by selecting the proper statistics, new learning algorithms which are not only convergent but rapid as well.

The effects of applying the viewpoints of stochastic approximation to nonsupervised learning become even more pronounced when most of the existing learning schemes can be considered, in one way or another, as to estimating the unknown parameters in a mixture distribution. By successively obtaining the estimates through the moment estimators of the mixture distribution, stochastic approximation algorithms are found to be effective tools in achieving a unique solution for the unknown parameters.

Finally, in addition to the results noted, a significant conclusion is that the modeling of both supervised and nonsupervised learning in terms of the stochastic approximation algorithms has resulted in a unified mathematical foundation for the analysis and synthesis of learning system in many applications.

APPENDIX

Dvoretzky's theorem: Suppose X_1 and Y_n assume values in normed space N with $||r||$ denoting the norm of r. Let θ be an element of N and $T_n(r_1, r_2, \ldots, r_n)$ be the measurable transformations from the nth Cartesian power of N into N and assume

(A1) $||T_n(r_1, r_2, \ldots, r_n)-\theta|| \le F_n \, ||r_n-\theta||$ with $F_n > 0$, $n=1,2,\ldots$, a sequence of positive numbers satisfying

(A2) $\prod_{n=1}^{\infty} F_n = o$. Then, if we put

(A3) $X_{n+1} = T_n(X_1, X_2, \ldots, X_n) + Y_n$, the assumptions

(A4) $E\{||X_1||^2\} + \sum_{n=1}^{\infty} \{||Y_n||^2\} < \infty$ and

(A5) $E\{||\varphi(X_1, X_2, \ldots, X_n) + Y_n||^2\} \le E\{||\varphi(X_1, X_2, \ldots, X_n)||^2\} + E\{||Y_n||^2\}$ for every measurable function $\varphi(X_1, X_2, \ldots, X_n)$ imply

(A6) $\lim_{n \to \infty} E\{||X_n-\theta||^2\} = o$ and

(A7) $P\{\lim_{n \to \infty} ||X_n-\theta|| = o\} = 1$.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Z. J. Nikolic for the helpful discussions.

The work was supported by the National Science Foundation Grant GK-696.

REFERENCES

[1] Abramson, N., and D. Braverman, "Learning to Recognize Patterns in a Random Environment," IRE Trans. on Information Theory, vol. IT-8, pp. S-58 to S-63, September 1962.

[2] Keehn, D.G., "A Note on Learning for Gaussian Properties", IEEE Trans. on Information Theory, vol. IT-11, pp. 126-132, January 1965.

[3] Daly, R.F., "The Adaptive Binary-Detection Problem on the Real Line", Stanford Electronics Labs., Stanford, California, Tech. Report 2003-3, February 1962.

[4] Fralick, S.C., "Learning to Recognize Patterns without a Teacher", Stanford Electronics Labs., Stanford, California, Tech. Report 6103-10, March 1965.

[5] Cooper, D.B., and P.W. Cooper, "Nonsupervised Adaptive Signal Detection and Pattern Recognition", Information and Control, vol. 7, No. 3, September 1964.

[6] Patrick, E.A., and J.C. Hancock, "Nonsupervised Sequential Classification and Recognition of Patterns", IEEE Trans. on Information Theory, vol. IT-12, pp. 362-372, July 1966.

[7] Teicher, H., "On the Mixture Distributions", Ann. Math. Stat., vol. 31, pp. 55-73, 1960.

[8] Robbins, H., and S. Monro, "A Stochastic Approximation Method", Ann. Math. Stat., vol. 22, pp. 400-407, 1951.

[9] Dvoretzky, A., "On Stochastic Approximation, "Proc. of the third Berkeley Symposium on Math. Stat. and Prob., University of California Press, Los Angeles, vol. 1, pp. 39-55, 1956.

[10] Fu, K.S., et. al., "On the Stochastic Approximation and Related Learning Techniques", TR-EE-66-6, School of Electrical Engineering, Purdue University, April 1966.

THE SYNTHESIS OF CASCADE LATTICE CRYSTAL FILTERS

JAMES D. SCHOEFFLER
Case Institute of Technology
Cleveland, Ohio 44106

ALLAN D. WAREN
Clevite Corporation
Cleveland, Ohio

ABSTRACT

A technique is presented for transforming a ladder network of crystal resonators into a cascade connection of symmetric lattices, each arm of which is a crystal resonator. The lattices are connected by capacitance L-sections. If the resonant and antiresonant frequencies of the resonators in the shunt arms of the ladder are larger than the corresponding frequencies in the series arms, the resulting cascade lattice filter necessarily has all positive elements and no transformers. Since this is the usual situation for band pass crystal ladder filters, a large class of insertion loss functions may be realized in the form of a cascade lattice structure. In addition, the two arms of each lattice have resonators with equal percent separation of the resonant and antiresonant frequencies.

Methods for transforming the capacitance interstage networks (either to eliminate them or else to control the resonant and antiresonant frequencies of the lattice resonators) are presented including one based on an optimization procedure which lends itself to efficient computer solution. Example designs including transformation of the interstage network are presented and evaluated.

1. INTRODUCTION

A large class of practical insertion loss functions can be synthesized in the form of a ladder network of resonators, terminated at each end with a capacitance and load resistance.[1, 2] Such a design is usually carried out on a digital computer and a special transformation of variables is used to minimize the inherent computational inaccuracy.[3, 4, 5] The ladder network however, has several drawbacks; the main one being the wide range of crystal resonator parameter values which is required. A more desirable structure is the cascaded symmetric lattice network which realizes its transmission zeros by cancellation, rather than by branch impedance poles or zeros as does the ladder network. General synthesis procedures, comparable to those for ladder resonator filters, are not available for this class of networks.

The objective of this paper is to present a synthesis technique for a class of resonator-realizable cascaded symmetric lattice filters and to derive conditions under which the desired network exists. It is shown that under almost all conditions, any band pass insertion loss function realizable as a resonator ladder network is also realizable as a cascade connection of symmetrical lattice networks, with one resonator per arm, together with an L-section of capacitance between each pair of lattices. Conditions under which the interstage capacitors may be totally absorbed into the lattices are also derived.

However, if it is not essential that all the inter-stage capacitance be absorbed, partial capacitance absorption can be used to obtain an additional degree of freedom. Since, as is true with all synthesis procedures, the spread of element values cannot be controlled during the synthesis, the resulting network is used as a starting point for a computer optimization program. This program distributes the capacitance throughout the network, without changing the insertion loss, in such a way as to optimize a performance function. Any measure of element value, element spread, or amount of interstage capacitance can be used as the performance function.

2. LADDER TO CASCADE LATTICE TRANSFORMATION

The approach adopted in this paper, to design crystal resonator cascade lattice filters, is to transform a resonator ladder band pass filter to the desired form. Similar transformation methods have been suggested previously for special cases or without realizability conditions.[6-11] The method presented here is exact, guarantees realizability, and is computationally straightforward.

Necessary and sufficient conditions for an insertion loss function to be realizable as a resonator ladder band pass filter are well known.[1] The usual results for such a realization are that series resonator impedance poles (transmission zeros) lie in the upper stop band and impedance zeros are in the pass band, whereas shunt resonator impedance zeros (also transmission zeros) lie in the lower stop band and poles in the pass band. Although these results are not guaranteed, they have proved true for all filters designed by the method to date. Thus, for these networks, impedance zeros and poles of series resonators are higher in frequency than those of shunt resonators.

Theorem I

A two resonator L-section is exactly equivalent to a symmetric lattice network, with one resonator per arm, cascaded with two capacitances and an ideal transformer (with all elements positive) if the pole and zero of the series resonator are higher in frequency than those of the shunt resonator. The equivalent networks are shown in Figure 1.

Proof

The equivalent circuit of the L-section, together with a pole zero plot of the admittance of each resonator, is shown in Figure 2. The relative locations of poles and zeros of the two arms obey the premises of Theorem I. Total removal, as a series capacitance, of the series resonator impedance pole at the origin causes the impedance zero at w_a to move to the origin. Partial removal of this impedance pole causes w_a to shift only part way toward the origin and, since $w_a > w_c$, it can be made to coincide with w_c. The result is the removal of a positive series capacitance from y_a leaving the remainder y_a' with a pole-zero plot shown in Figure 2.

The shunt arm admittance has a pole at infinity. Total removal of this pole, as a shunt capacitor, causes the admittance zero w_d to move to infinity. Since $w_d < w_b$, partial removal of the admittance pole at infinity, which shifts w_d only part way toward infinity, can force w_d to coincide with w_b. This results in the removal of a shunt capacitance from y_b, leaving a remainder y_b' with the pole-zero plot shown in Figure 2.

Removal of a series and a shunt capacitance, at the input and output sides of the L-section respectively, leaves the two arms of the L-section with equal resonant and antiresonant frequencies. Because the two capacitances removed corresponded to partial pole removals, both admittances have zeros at the origin and poles at infinity. Consequently they differ only in their residue of the poles at infinity. Denote the two admittances as

$$y'_a = \frac{K_a s(s^2 + w_b^2)}{s^2 + w_c^2}$$

$$y'_b = \frac{K_b s(s^2 + w_b^2)}{s^2 + w_c^2} \tag{2.1}$$

The short-circuit admittance matrix of the L-section two port may then be written as

$$[y] = \frac{s(s^2 + w_b^2)}{(s^2 + w_c^2)} \begin{bmatrix} K_a & -K_a \\ -K_a & K_a + K_b \end{bmatrix} \tag{2.2}$$

If an ideal transformer is removed from the secondary, with the proper turns ratio, the above admittance matrix may be made symmetric. Choose a turns ratio

$$n = \sqrt{\frac{K_a}{K_a + K_b}} \tag{2.3}$$

Clearly n is positive and less than unity. This transformer removal causes the admittance matrix to become (second row and column are multiplied by n)

$$[y'] = \frac{s(s^2 + w_b^2)}{s^2 + w_c^2} \begin{bmatrix} K_a & -K_a n \\ -K_a n & K_a \end{bmatrix} . \tag{2.4}$$

$[y']$ corresponds to a symmetric lattice network with arm admittances y_1 and y_2 given by

$$y_1 = \frac{K_a(1-n)\, s(s^2 + w_b^2)}{s^2 + w_c^2}$$

$$y_2 = \frac{K_a(1+n)\, s(s^2 + w_b^2)}{s^2 + w_c^2} \tag{2.5}$$

Thus, given an L-section of resonators, with the specified pole-zero frequency relations, it may be transformed to a symmetric lattice network with one resonator per arm, with a series capacitance on the input side and a transformer and shunt capacitance on the output side. Moreover, all element values are guaranteed positive. This completes the proof of Theorem I.

Using Theorem I, the following important result may be derived:

Theorem II

Any crystal resonator ladder filter, such that series (shunt) resonators have resonances (antiresonances) in the pass band and antiresonances (resonances) in the upper (lower) stop band may be transformed to a cascade connection of symmetric lattices, with one resonator per arm, coupled by L-sections of capacitance with all elements positive.

Proof

The general form of a ladder network of resonators is as shown in Figure 3. (For a derivation of this network form and necessary and sufficient conditions under which it exists, see Reference 1.) Divide the network into a cascade connection of L-sections of resonators (if there are an odd number of resonators, one resonator must be shared by two L-sections). Each L-section must satisfy the condition of Theorem I by hypothesis and consequently may be transformed to a symmetric lattice network with a series capacitance at its input and a transformer and shunt capacitance at its output. The transformers may be removed, without changing the insertion loss of the filter, merely by impedance leveling all of the elements to the right or left of the transformer. The result is an L-section of capacitance between each pair of symmetric lattices, as shown in Figure 4, and all elements are still positive.

The ladder network terminations consist of a series or shunt capacitance together with a load or input resistance. These termination networks are also impedance scaled in the removal of the transformers and the series or shunt capacitance of the L-section adjacent to the termination is added. In general, the result will be a resistor in series with a capacitance and shunted by a second capacitance, the same form as a resonator except that the inductance has been replaced by a resistance. This final network form is also shown in Figure 4. All elements are positive and there are no transformers in the network. This completes the proof of Theorem II.

One characteristic of the resulting network is particularly important: the locations of the frequencies of the lattice resonators. Define the percent separation between the resonant and antiresonant frequencies of a resonator as

$$\Delta f/f_r = \frac{f_a - f_r}{f_r} \tag{2.6}$$

where f_a and f_r are the antiresonant and resonant frequencies respectively. In practice, this ratio is limited as to maximum and minimum size and it is inconvenient in manufacturing to have quite different ratios in the filter. Note that y_a' and y_b' are identical except for gain because of the way capacitance was removed. Thus their resonant and antiresonant frequencies are the same and the following observation becomes obvious:

Observation

In each symmetric lattice, all the resonators have the same $\Delta f/f_r$ ratios. It is of interest to determine the relationship between this ratio and the ratios of the original ladder resonators.

Note in Figure 2 that if the resonant frequency of the series arm and the antiresonant frequency of the shunt arm occur at the same frequency in the pass band, the total separation between resonant and antiresonant frequencies is the sum of the separations of the original ladder network resonators. Since the resonant frequencies differ only slightly, the final $\Delta f/f_r$ ratio is the sum of the ratios of the original resonators. If the two pass band frequencies are not coincident, they can differ at most by the width of the pass band. Clearly then the separation between the resonant and antiresonant frequencies of the lattice resonators is the sum of the separations of the original ladder resonators plus or minus the bandwidth of the filter. Usually the ratio is close to the sum of the original ratios.

From a practical point of view, the equal $\Delta f/f_r$ ratios is desirable since fewer different resonator types must be used and the capacitance coupling networks is a small price to pay. On the other hand, it may be desirable to absorb the interstage networks into the lattices thereby decreasing the $\Delta f/f_r$ ratios and reducing the total number of elements. Methods for accomplishing this end are discussed in the next section.

3. ABSORPTION OF CAPACITANCE INTERSTAGE NETWORKS

In many applications, the presence of the capacitance interstage networks may not be objectionable. For these cases it would be desirable to control the capacitance within the lattices so that the lattices could be realized with resonators with as few padding capacitors as possible. In other applications, where the interstage networks are undesirable, it would be advantageous to absorb them into the lattices, if possible.

Two simple lattice equivalences are illustrated in Figure 5. Using these equivalences, together with impedance leveling, capacitance may be added or removed from the lattices thus providing great control over element values. However, the order in which series and parallel capacitance should be added and removed in a large network, in order to achieve some optimal performance (element value distribution, number of elements, etc.), is not clear. Since there are such a large number of possibilities it would be desirable to formulate this transformation problem in such a way that all these possibilities need not be individually investigated.

To this end, observe that any application of the equivalence shown in Figure 5 results in the placing of a capacitance in series or parallel with the arms of the lattices or interstage networks. Repeated application of these equivalences, plus impedance leveling, results in each resonator, termination network, or interstage network being cascaded with a ladder network of capacitances and ideal transformers. Considering each such ladder to be a two-port, one-element kind network, there exists a capacitive Pi network equivalent to it. This fact leads to the following result:

Theorem III

Any sequence of steps of series and shunt capacitance absorption, plus impedance leveling, is equivalent to the addition of a Pi network of capacitance as shown in Figure 6.

This result can be used advantageously by observing the circuit shown in Figure 7. The circuit formed by interchanging the input and output terminals, and replacing each element by its negative, has an ABCD parameter matrix which is the inverse of that for the original network. Thus the cascade connection of such a network with its original is equivalent to a direct connection from input to output. Capacitive Pi networks of this form will be denoted as shown in Figure 8 for convenience. Hence the cascade connection of a "positive" and a "negative" Pi network anywhere in a circuit does not affect the transmission functions of the network.

Using the above results, the cascade lattice network can be redrawn as shown in Figure 9. The positive Pi networks correspond to any optimal sequence of capacitance absorption and impedance leveling and the negative Pi networks are present to assure no change in the insertion loss. The positive Pi networks may be shifted into the lattices, resulting in a Pi network of capacitance terminated in a resonator for each lattice arm. Each end of the overall network is terminated in a similar capacitive Pi terminated with an R. C. network. Of course, the capacitive Pi terminated with a resonator is equivalent to a resonator with different parameter values (since the resonator already has an impedance pole at the origin and an admittance pole at infinity, the capacitance network merely modifies the residues at these poles). The negative Pi networks are combined with the interstage networks and, if properly chosen, result in either elimination of the interstage networks or optimal performance in some sense. Hence if the parameters of the Pi networks are selected so that realizable terminating networks, realizable resonators, and realizable interstage networks result, a realizable equivalent network results.

In order to formulate the restrictions imposed on the elements of the Pi networks so that the resulting resonators and terminating networks are realizable, it is convenient to use the ABCD matrix formulation. The Pi network of capacitance shown in Figure 10 has the following ABCD matrix:

$$a = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = 1/C_b \begin{bmatrix} C_b+C_c & 1/s \\ s(C_aC_b+C_aC_c+C_bC_c) & C_a+C_b \end{bmatrix} . \qquad (3.1)$$

Note that in this formulation, the current I_2 is assumed to be positive when flowing out of the two port rather than into the two port. The capacitances of the Pi network are simply related to the ABCD parameters:

$$\begin{aligned} C_a &= (D-1)/Bs \\ C_b &= 1/Bs \\ C_c &= (A-1)/Bs \quad . \end{aligned} \qquad (3.2)$$

Clearly, necessary and sufficient conditions under which an ABCD matrix possesses a Pi equivalent with all positive capacitances are that

$$\begin{aligned} Bs &\geq 0 \\ A &\geq 1 \\ D &\geq 1 \end{aligned} \qquad (3.3)$$

Consider the Pi network terminated in either an RC network (this is pertinent to the termination networks) or a resonator network as illustrated in Figure 11. The R(or L) termination has been replaced by a scaled R (or L) together with a transformer. This impedance leveling permits reduction of the Pi network to an L-section. If the Pi network has ABCD matrix with elements z_1, z_2, z_3, z_4 and the two capacitances are denoted by x and y as shown in Figure 11, then (assuming x and y positive), necessary and sufficient conditions under which x' and y' are positive are

$$\begin{aligned} n &= z_1 + z_2 y \geq 0 \\ n &+ z_2 + \geq 0 \\ z_3 &+ z_4 y \geq 0 \end{aligned} \qquad (3.4)$$

where all three inequalities are in the same direction. Here it is recognized that the determinant of the ABCD matrix, $z_1 z_4 - z_2 z_3$, must be unity.

The values for x' and y' and n are:

$$n = z_1 + z_2 y$$

$$y' = (nz_4 - 1)/nx_2$$

$$1/x' = (n(z_2 + n/x))$$

$$R' = n^2 R(L' = n^2 L) \tag{3.5}$$

The proof is very straightforward and involves simply writing out the ABCD matrix of the cascade connection and forcing the three realizability conditions to hold. The steps are presented here for completeness.

The ABCD matrix of the capacitance L-section in Figure 11 is simply (see Figure 12)

$$\begin{pmatrix} E_1 \\ I_1 \end{pmatrix} = \begin{pmatrix} 1 & 1/xs \\ sy & (x+y)/x \end{pmatrix} \begin{pmatrix} E_2 \\ I_2 \end{pmatrix} \tag{3.6}$$

The transformer has the ABCD relations

$$\begin{pmatrix} E_2 \\ I_2 \end{pmatrix} = \begin{pmatrix} 1/n & 0 \\ 0 & n \end{pmatrix} \begin{pmatrix} E_3 \\ I_3 \end{pmatrix} \tag{3.7}$$

The cascade connection of the Pi network, the L-section, and the transformer must be equivalent to an L-section. Thus we require

$$\begin{pmatrix} Z_1 & Z_2 \\ Z_3 & Z_4 \end{pmatrix} \begin{pmatrix} 1/n & n/xs \\ sy/n & n(1+y/s) \end{pmatrix} = \begin{pmatrix} 1 & 1/x's \\ sy' & 1+y'/x' \end{pmatrix} \tag{3.8}$$

Equating corresponding elements of the ABCD matrix results in the equalities:

$$z_1 = z_2 y = n$$

$$n[z_1/x + z_2(1+y/x)] = 1/x'$$

$$(z_3 + z_4 y)/n = y' \tag{3.9}$$

Thus if n is positive, we require that all of the other terms be positive and vice versa. In other words

$$\begin{aligned} n &\geq 0 \\ n + z_2 x &\geq 0 \\ z_3 + z_4 y &\geq 0 \end{aligned} \tag{3.10}$$

where all three inequalities are in the same direction.

This result may be applied to each resonator and each terminating network in order to determine the restrictions on the choice of the Pi elements.

For clarity, the procedure is illustrated for the two lattice case shown, together with the Pi networks, in Figure 13. Figures 14a through 14e illustrate the identification of the appropriate ABCD matrix with the Z-matrix and the corresponding constraints which result. There are three sets of constraints on the parameters of each of the two Pi networks. Figure 14e also shows the calculation yielding the ABCD matrix of the interstage network and the values for the three capacitances of this network. If the problem is to minimize the total interstage capacitance subject to realizability of the network, it is necessary to select the 6 elements of the two Pi networks such that all of the constraints are satisfied and, in addition, the three capacitances of the interstage Pi are positive and have a minimum value for their sum. Clearly any other criterion could also be applied. The formulation and solution of the optimization problem is discussed in the next section.

4. FORMULATION OF THE OPTIMIZATION PROBLEM

Depending upon the particular application at hand, several goals for the control of the interstage network and lattice capacitances are appropriate. As an example, consider the elimination of the interstage networks. Define a performance function $f(x)$ equal to the sum of the shunt capacitances in the interstage networks plus the reciprocals of the series capacitances. The vector $\underline{x}$ is the vector of manipulated variables, the elements of the arbitrary $\overline{\text{Pi}}$ networks introduced in Figure 13. The elements of $\underline{x}$ are not entirely arbitrary since the resulting lattice resonators are constrained to remain physically realizable and if the interstage capacitances do not vanish, they too must remain positive. Denote the resulting constraint equations by

$$g_i(\underline{x}) \geq 0 \qquad i = 1, 2, \ldots, N \tag{4.1}$$

where N is the number of constraints (15 in the example problem: 2 for each resonator, 2 for each termination network and one for each capacitive interstage network).

Previous work[12] has indicated that this constrained optimization problem may be solved most efficiently by the Fialkow-McCormick method which forms the unconstrained optimization problem:

$$\underset{\underline{x}}{\text{Min}} \;\; \phi(\underline{x}, r) = f(\underline{x}) + r \sum_{i=1}^{N} (g_i(\underline{x}))^{-1} \tag{4.2}$$

$\phi(\underline{x}, r)$ is then minimized for a decreasing sequence of values of r. In the limit as r approaches zero, the constrained minimum is found. In practice, only a few values of r have been found to be necessary and the overall solution is efficient since the starting point for successive values of $\underline{r}$ is improved considerably with each iteration.

Minimization of $\phi(x, r)$ for a given value of r is an unconstrained minimization problem. Experience indicates that this problem is handled most efficiently by the Fletcher-Powell method when gradients of the function ϕ can be efficiently calculated.[12] The computation of the gradient of $\phi(\underline{x}, r)$ with respect to $\underline{x}$ is a straightforward procedure in the problem at hand. Each of the constraint equations may be written explicitly in terms of the manipulated variables. Since each of the interstage capacitances is given explicitly in terms of the manipulated variables also in Figure 13, computation of the necessary gradient may be done explicitly and is sufficiently efficient. Note that the gradient complexity increases only linearly with the number of resonators since each Pi network (manipulated variables) affects at most two interstage capacitance networks. The major core storage requirement in the Fletcher-Powell method is a square matrix whose dimension is equal to the number of manipulated variables. This number is 3 per lattice or 3/4 of the number of resonators and is not excessive even for the largest filters of interest.

As an example of this design process, a resonator-ladder network realization of an insertion loss function having equi-ripple character in both the pass and stop bands was derived using the program described in reference 12. The network is shown (normalized to a center frequency of 1 radian/second) in Figure 5. The insertion loss realized by this filter has a maximum pass band attenuation of 1.0 db and a minimum stop band attenuation of 37.1db (assuming lossless resonators). The filter has four frequencies of infinite attenuation, two in the lower and two in the upper stop bands. The result of the transformation from the ladder to the cascaded lattice network with capacitance interstage networks is shown in Figure 15. The interstage network in this filter can be totally removed by three applications of the network equivalences in Figure 4. The resulting cascaded lattice network is shown in Figure 16. To compare results, the optimization program was applied to the filter with interstage network (Figure 15). After four values of r, the filter shown in Figure 17 resulted. Notice that the capacitance has been distributed quite differently from that in Figure 16 but that the interstage network has again been essentially removed. This extra degree of freedom suggests that is is appropriate to include more than merely the interstage capacitance in the optimization criterion. In addition, the capacitance ratios of the resonators together with the percent separation of the resonant and antiresonant frequencies should be controlled. Nontheless, the power of this approach has been demonstrated, namely, the use of analytical design methods to obtain a good starting point for the computer optimization program is an efficient approach to network design. In the example problem here, total time to compile the program (Algol 60), transform the ladder to the cascade lattice and complete the optimization iteratively was 103 seconds (Univac 1107).

REFERENCES

1. J. D. Schoeffler, "On the Existence and Realization of Crystal Ladder Filters," Proc. of Allerton Conf. on Circuit and System Theory, Vol. 1, pp. 259-274, November, 1963.

2. J. D. Schoeffler, "A Solution to the Approximation and Realization Problems for Crystal Ladder Filters," IEEE Conv. Record, Part 1, pp. 282-288, March, 1964.

3. G. Szentirmai, "Theoretical Basis of a Digital Computer Program Package for Filter Synthesis," Proc. University of Illinois, First Allerton Conf. on Circuit and System Theory, pp. 37-49, 1963.

4. G. Szentirmai, "The Design of Arithmetically Symmetrical Band-Pass Filters," IEEE Trans, Vol. CT-10, No. 3, Sept., 1963.

5. J. D. Schoeffler and A. D. Waren, "Approximation Problem for Resonator Ladder Filters," IEEE Trans., Vol. CT-12, No. 2, June, 1965 pp. 215-222.

6. T. R. O'Meara, "The Cascade Synthesis of Crystal Filters with Transmission Zeros at Infinity," IEEE Trans., Vol. CT-10, No. 4, Dec., 1963, pp. 533-534.

7. Editors, "Network Synthesis of Narrow-Band Crystal Filters," Electro-Technology, November, 1965, pp. 36-43.

8. R. C. Smythe, "On the Realization of Crystal-Capacitor Tanden Lattice Band-Pass Filters," IEEE Trans., Vol. CT-11, No. 1, March, 1964, pp. 170-171.

9. G. Szentirmai, "On the Realization of Crystal Band-Pass Filters," IEEE Trans., Vol. CT-11, No. 2, June, 1964, pp. 299-301.

10. J. A. C. Bingham, "A Calculation Method for Szentirmai's Modified Symmetrical Lattice Circuit," IEEE Trans. Vol. CT-12, No. 2, June, 1965, pp. 284-285.

11. J. Long, C. E. Schmidt, "Crystal Filter Transformations," IEEE Trans., Vol. CT-12, No. 3, Sept., 1965, pp. 454-457.

12. A. D. Waren and L. S. Lasdon, "Practical Filter Design Using Mathematical Optimization," Proc. University of Illinois, Third Allerton Conf. on Circuit and System Theory, pp. 677-689, 1965.

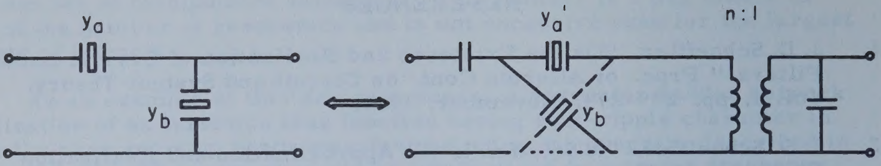

y_a
y_b
y_a'
y_b'
n:1

FIG. 1. EQUIVALENT CIRCUITS

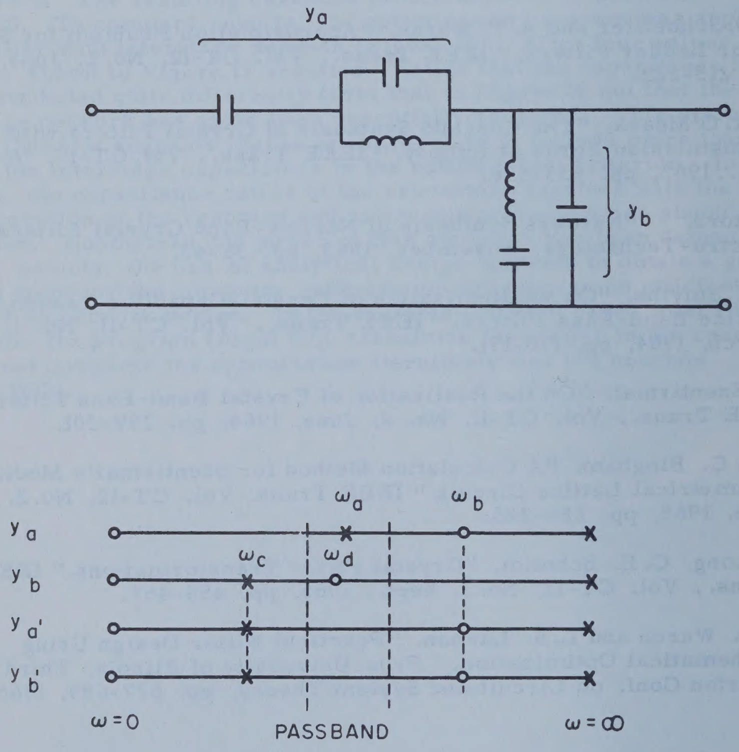

y_a
y_b
ω_a
ω_b
y_a
ω_c
ω_d
y_b
y_a'
y_b'
$\omega = 0$
PASSBAND
$\omega = \infty$

FIG. 2. L-SECTION EQUIVALENT CIRCUIT

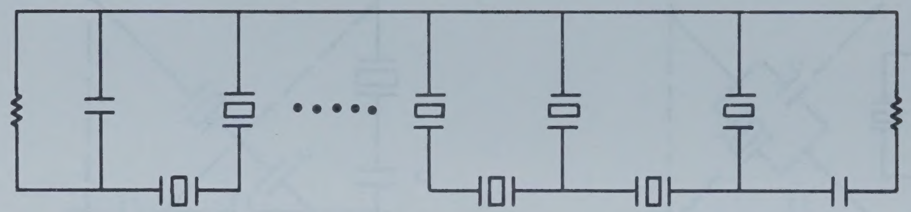


FIG. 3. LADDER NETWORK OF CRYSTAL RESONATORS

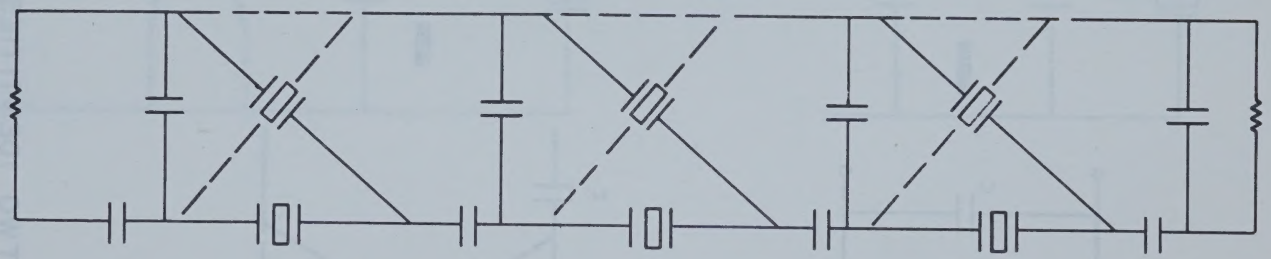


FIG. 4. CASCADE LATTICE EQUIVALENT CIRCUIT

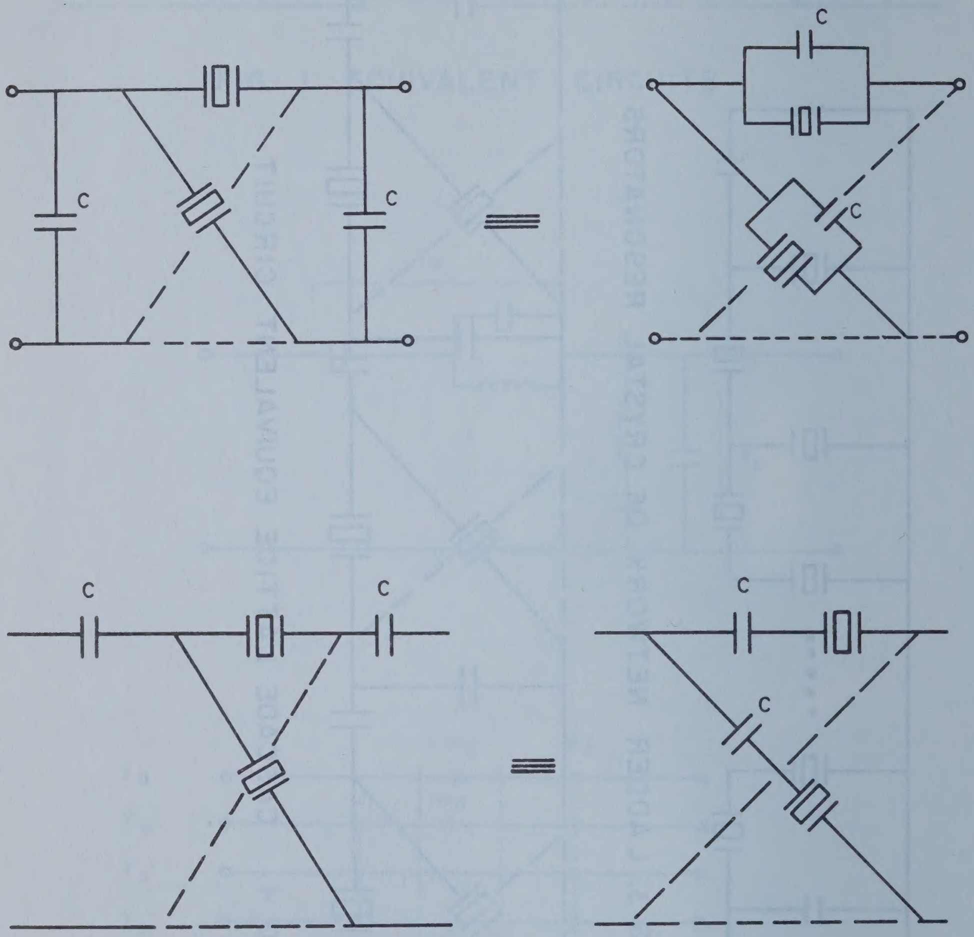

C
C
C
≡
C
C
C
C
C
≡
C
C

FIG. 5. TWO IDENTITIES

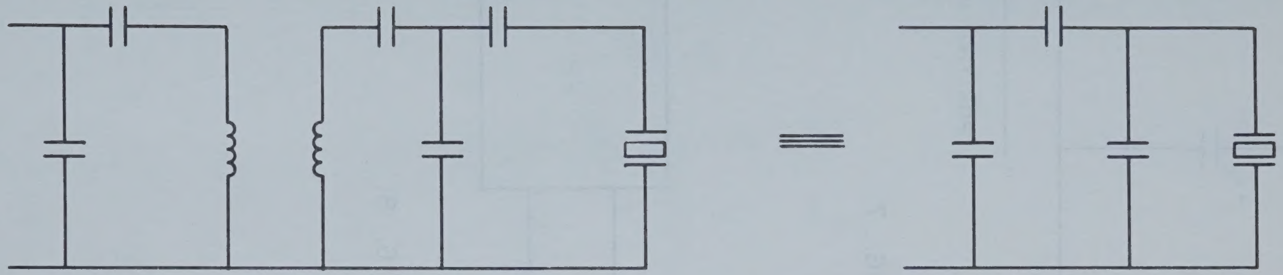
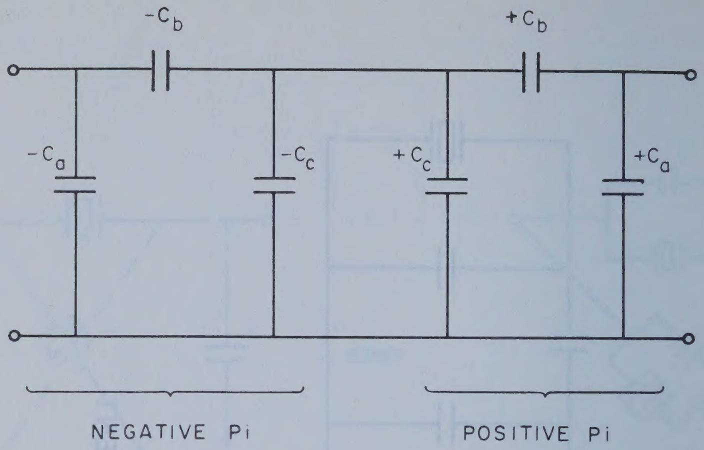


FIG. 6. TRANSFORMER-CAPACITOR LADDER EQUIVALENT



$-C_b$
$+C_b$
$-C_a$
$-C_c$
$+C_c$
$+C_a$
NEGATIVE Pi
POSITIVE Pi

FIG. 7.

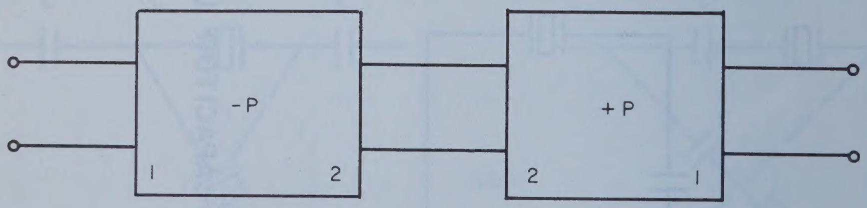


FIG. 8.

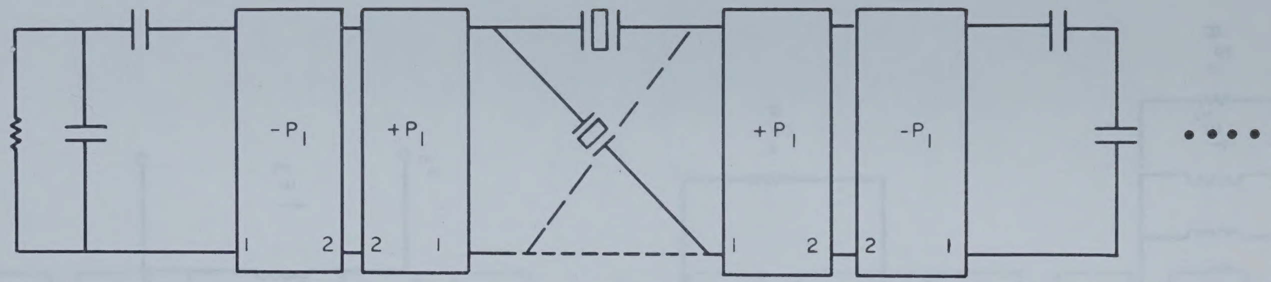
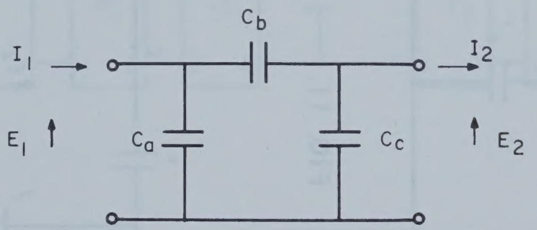

-P1
+P1
+P1
-P1
1
2
2
1
1
2
2
1

FIG. 9.


C_b
I_1
I_2
E_1
C_a
C_c
E_2

$$E_1 = \left(\frac{C_b + C_c}{C_b}\right) E_2 + \left(\frac{1}{SC_b}\right) I_2$$

$$I_1 = S\left(\frac{C_aC_b + C_aC_c + C_bC_c}{C_b}\right) E_2 + \left(\frac{C_a + C_b}{C_b}\right) I_2$$

FIG. 10. CAPACITANCE Pi NETWORK

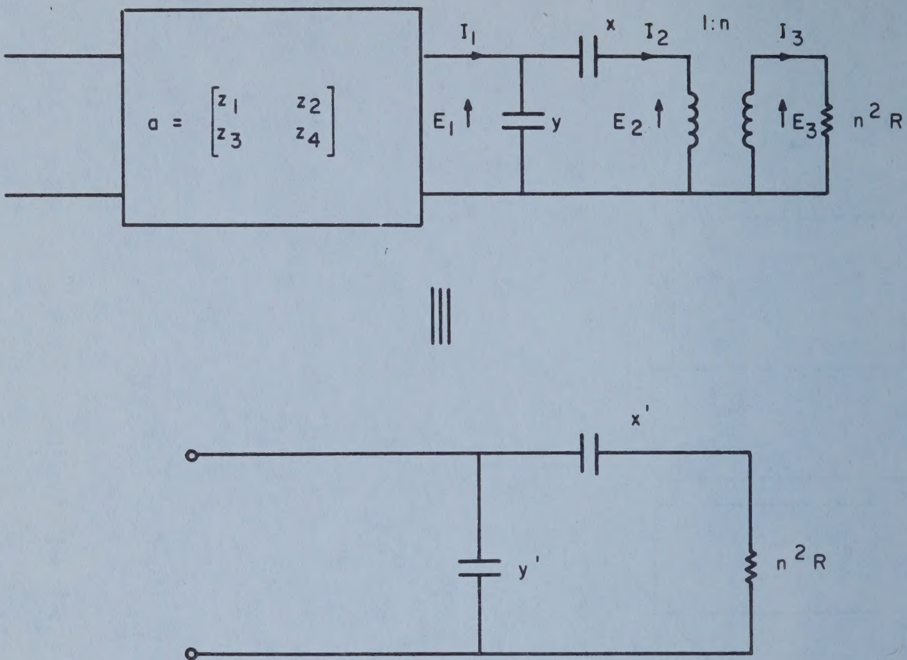

a = $\begin{bmatrix} z_1 & z_2 \\ z_3 & z_4 \end{bmatrix}$
I_1
x
I_2
1:n
I_3
E_1
y
E_2
E_3
n^2R
≡
x'
y'
n^2R

FIG. 11.

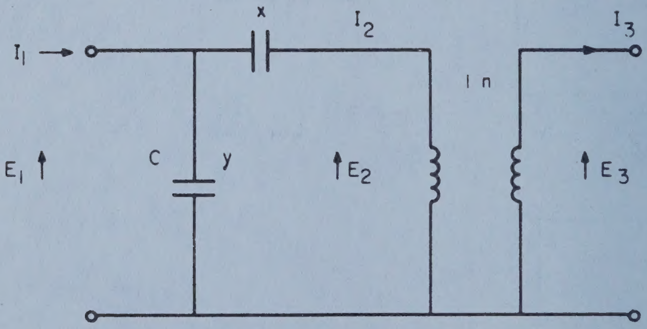

I_1
x
I_2
I_3
1 n
E_1
C
y
E_2
E_3

FIG. 12.

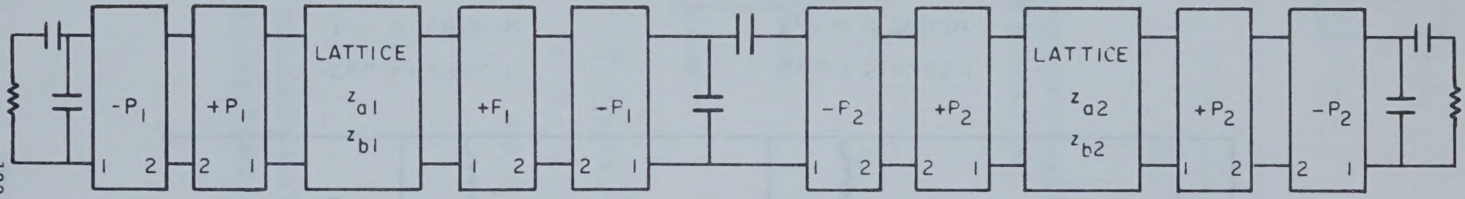

-P1
1
2
+P1
2
1
LATTICE
za1
zb1
+F1
1
2
-F1
2
1
-F2
1
2
+P2
2
1
LATTICE
za2
zb2
+P2
1
2
-P2
2
1

FIG. 13. TWO LATTICE FILTER CIRCUIT

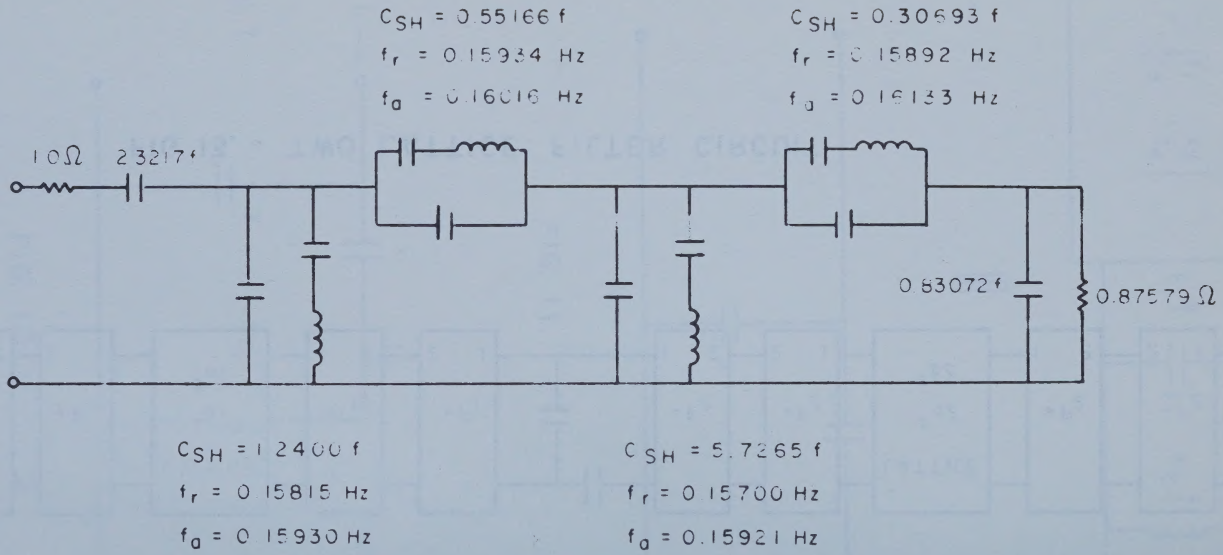
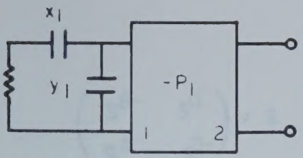

C_{SH} = 0.55166 f
f_r = 0.15934 Hz
f_a = 0.16016 Hz
C_{SH} = 0.30693 f
f_r = 0.15892 Hz
f_a = 0.16133 Hz
1 0Ω
2.3217 f
0.83072 f
0.87579 Ω
C_{SH} = 1.2400 f
f_r = 0.15815 Hz
f_a = 0.15930 Hz
C_{SH} = 5.7265 f
f_r = 0.15700 Hz
f_a = 0.15921 Hz

FIG. 14. INITIAL RESONATOR LADDER FILTER.



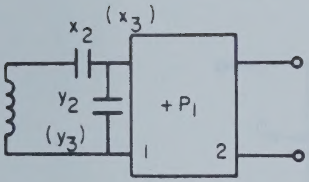
$$Z = \begin{pmatrix} D_1 & B_1 \\ C_1 & A_1 \end{pmatrix}$$

$$(D_1 + B_1 y_1)(D_1 + B_1 y_1 + B_1 x_1) \geq 0$$

$$(D_1 + B_1 y_1)(C_1 + A_1 y_1) \geq 0$$

$$A_1 D_1 - B_1 C_1 = 1$$

FIG. 14a. TRANSFORMATION OF THE FILTER RESONATORS.



$$Z = \begin{pmatrix} D_1 & -B_1 \\ -C_1 & A_1 \end{pmatrix}$$

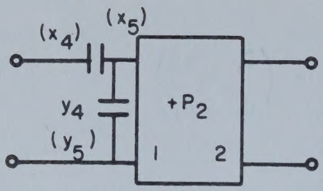
$$(D_1 - B_1 y_2)(D_1 - B_1 y_2 - B_1 x_2) \geq 0$$

$$(D_1 - B_1 y_2)(-C_1 + A_1 y_2) \geq 0$$

$$A_1 D_1 - B_1 C_1 = 1$$

SAME INEQUALITIES
FOR x_3 AND y_3

FIG. 14b. TRANSFORMATION OF THE FILTER RESONATORS.



$$Z = \begin{pmatrix} D_2 & -B_2 \\ -C_2 & A_2 \end{pmatrix}$$

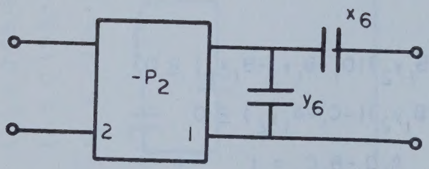
$$(D_2 - B_2 y_4)(D_2 - B_2 y_4 - B_2 x_4) \geq 0$$

$$(D_2 - B_2 y_4)(-C_2 + A_2 y_4) \geq 0$$

$$A_2 D_2 - B_2 C_2 = 1$$

SAME INEQUALITIES
FOR x_5 AND y_5

FIG. 14c. TRANSFORMATION OF THE FILTER RESONATORS.



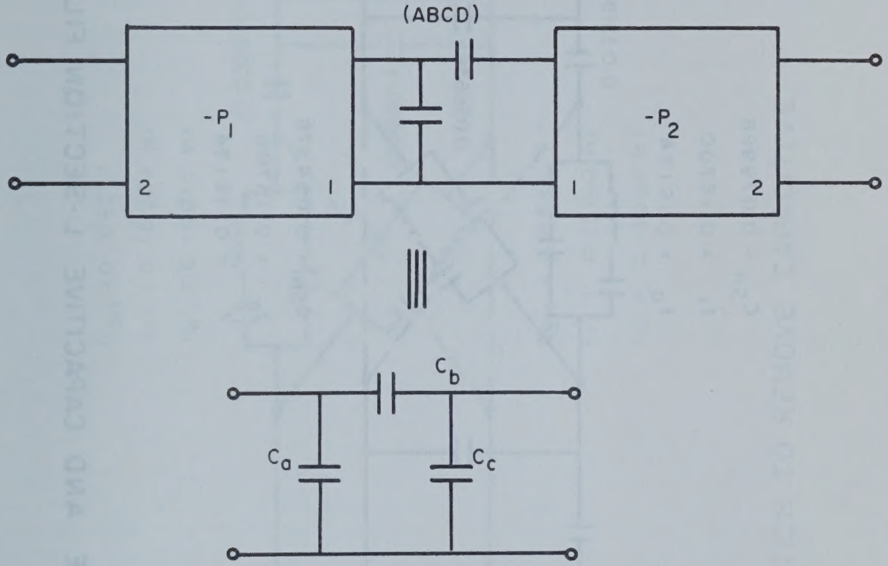
$$Z = \begin{pmatrix} D_2 & B_2 \\ C_2 & A_2 \end{pmatrix}$$

$$(D_2 + B_2 y_6)(D_2 + B_2 y_6 + B_2 x_6) \geq 0$$

$$(D_2 + B_2 y_6)(C_2 + A_2 y_6) \geq 0$$

$$A_2 D_2 - B_2 C_2 = 1$$

FIG. 14 d. TRANSFORMATION OF THE FILTER RESONATORS.



$$(A'B'C'D')_m = \begin{pmatrix} D_1 & B_1 \\ C_1 & A_1 \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} A_2 & -B_2 \\ -C_2 & D_2 \end{pmatrix}$$

$$C_a = (D'_m - 1)/B'_m \qquad C_b = 1/B'_m \qquad C_c = (A'_m - 1)/B'_m$$

FIG. 14e. TRANSFORMATION OF THE FILTER RESONATORS.

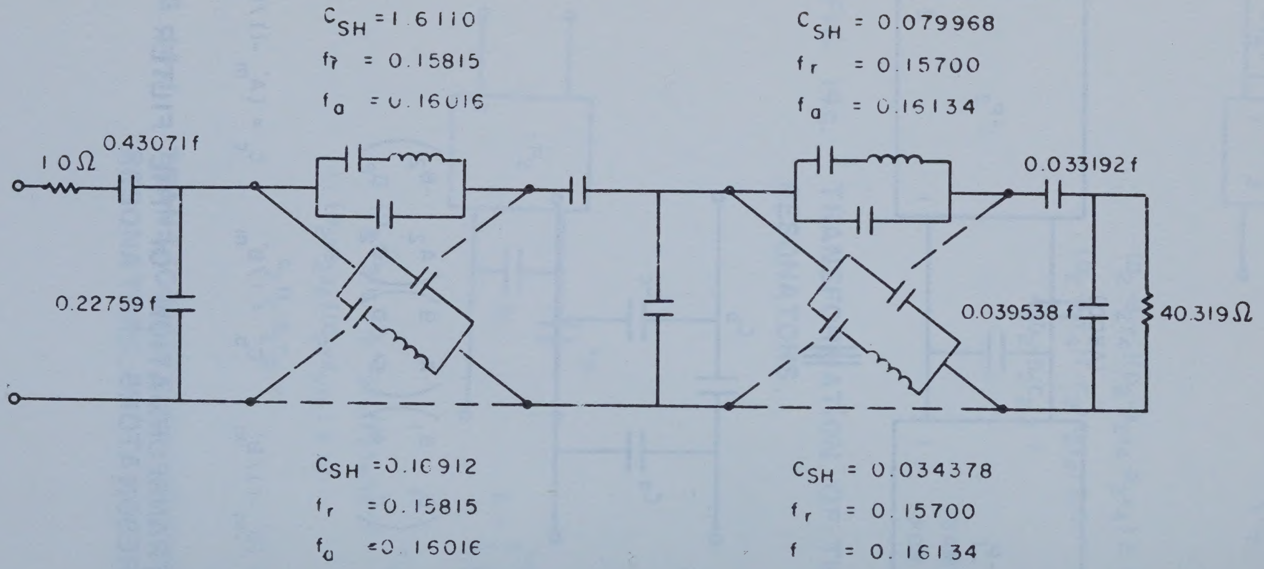
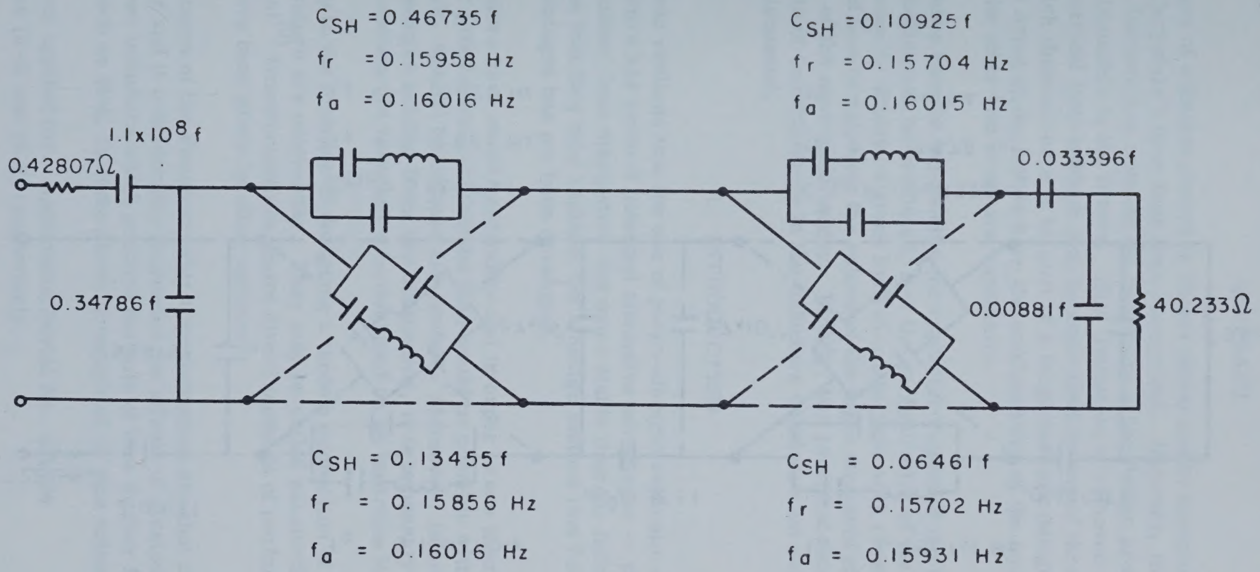

C_SH = 1.6110
fr = 0.15815
fa = 0.16016
C_SH = 0.079968
fr = 0.15700
fa = 0.16134
1 0Ω
0.43071f
0.22759f
0.033192f
0.039538f
40.319Ω
C_SH = 0.16912
fr = 0.15815
fa = 0.16016
C_SH = 0.034378
fr = 0.15700
f = 0.16134

FIG. 15. CASCADE LATTICE AND CAPACITIVE L-SECTION FILTER.



$C_{SH} = 0.46735$ f
$f_r = 0.15958$ Hz
$f_a = 0.16016$ Hz
$C_{SH} = 0.10925$ f
$f_r = 0.15704$ Hz
$f_a = 0.16015$ Hz
1.1×10^8 f
0.42807Ω
0.033396f
0.34786 f
0.00881f
40.233Ω
$C_{SH} = 0.13455$ f
$f_r = 0.15856$ Hz
$f_a = 0.16016$ Hz
$C_{SH} = 0.06461$ f
$f_r = 0.15702$ Hz
$f_a = 0.15931$ Hz

FIG. 16. DIRECT TRANSFORMATION TO REMOVE CAPACITIVE COUPLING SECTION.

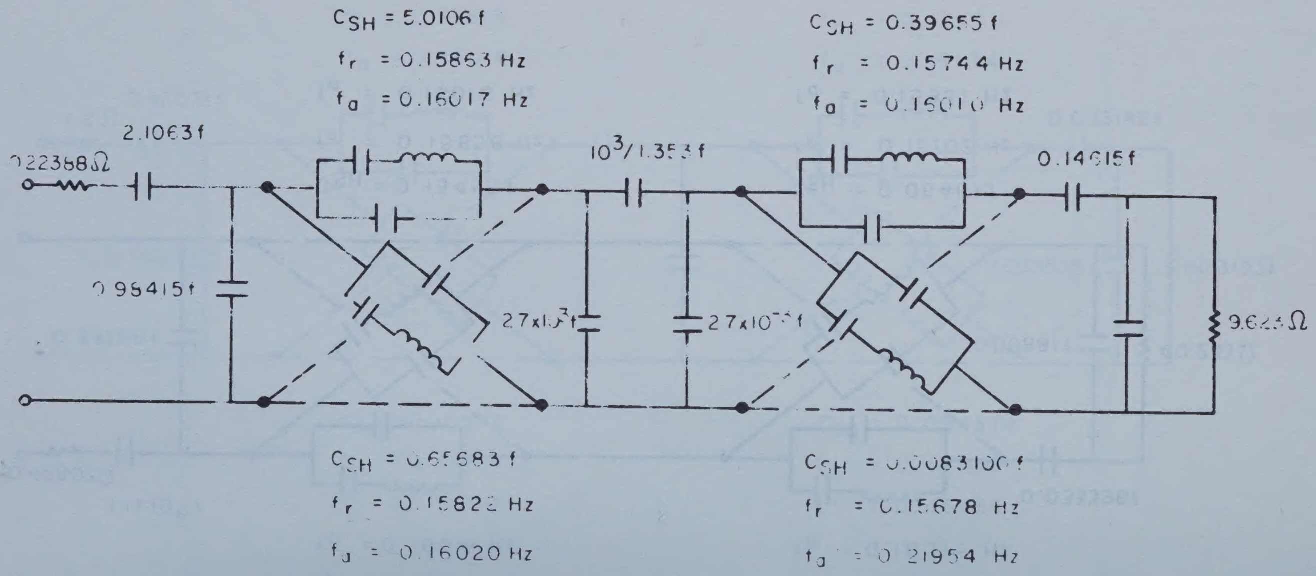

C_SH = 5.0106 f
f_r = 0.15863 Hz
f_a = 0.16017 Hz
C_SH = 0.39655 f
f_r = 0.15744 Hz
f_a = 0.16010 Hz
2.1063 f
.22388 Ω
10^3/1.353 f
0.14615 f
0.95415 f
27x10^-3 f
27x10^-3 f
9.623 Ω
C_SH = 0.65683 f
f_r = 0.15822 Hz
f_a = 0.16020 Hz
C_SH = 0.0083100 f
f_r = 0.15678 Hz
f_a = 0.21954 Hz

FIG. 17. CASCADE FILTER AFTER OPTIMIZATION

AN ASYMMETRICAL LATTICE[1] FOR CRYSTAL BAND-PASS FILTERS

JOHN A. C. BINGHAM
Lenkurt Electric Co., Inc.
San Carlos, California

ABSTRACT

The advantages of a lattice circuit in filters which are to incorporate piezo-electric resonators ("crystals") have long been recognized. However, the limitation to symmetrical lattices has reduced the flexibility of the design procedures and/or the bandwidths obtainable in the filters. This limitation is removed by the introduction of an asymmetrical lattice which may be equivalent to any of three small ladder networks which themselves may be part of a large network designed by any method. Two adjacent arms of the lattice have the configuration of the equivalent circuit of a crystal and the other two arms are capacitors.

The paper shows how the bandwidths (or capacitor ratios) of the two crystal-like arms of the lattice can be predicted, how the configuration of the overall filter should be chosen to achieve a given loss characteristic with crystals of a given bandwidth and how to calculate the values of the eight elements of the lattice from the elements of the equivalent ladder. Finally, two practical examples are shown and the effects of imperfections in the inductors which are not realized by crystals are briefly discussed.

1. INTRODUCTION

It has long been realized that the use of piezo-electric resonators in highly selective band-pass filters has several potential attractive advantages -- the resonators are generally smaller, less dissipative and more stable than the inductor/capacitor configurations that they may replace; but a design method that fully realizes these potential advantages has not been developed.

These resonators have small bandwidths and in order to use them in filters with much greater proportional bandwidths than would be possible with a conventional ladder network, special techniques are needed. Moreover, in order to realize fully the advantages arising from low dissipation, it is necessary that the filters which are to contain the resonators be designed on an insertion loss basis.

Two solutions to the problem of designing a tandem connection of networks which may contain resonators are noteworthy. They may be called succinctly the Poschenrieder(1) and Szentirmai(2) transformations; more direct methods of performing these transformations have been given by other authors[2].

The disadvantages of the Poschenrieder transformation are that it introduces an extra inductor and it considerably increases the effects of dissipation and instability in the two inductors which produce the pole of loss further from the pass band -- so much so that, in some cases investigated by this author, the effect of

1. U.S. Patent applied for. Application Serial No. 411034
2. References (3-5) and (6-9) respectively.

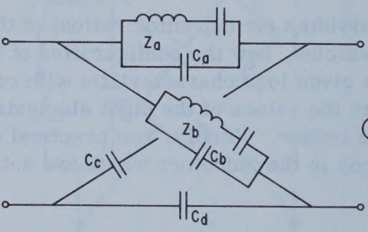
introducing a stable high-Q resonator only reduced the effects of dissipation in the transformed part of the ladder by ten percent. This effect has been discussed by Bohse(10).

The disadvantage of the Szentirmai transformation is that it very often leads to negative capacitance values in the case of lattices which produce two poles of loss in the same stop band[3].

A third solution is to abandon completely the restraints imposed by three-terminal networks and symmetrical networks and transform selected two-ports into asymmetrical lattices; this method is the subject of the present paper.

2. AN ASYMMETRICAL LATTICE

Consider the lattice shown in Fig. 1.



Let the impedances of the four arms be defined by

$$Z_a \triangleq \frac{p^2 + \omega_{a1}^2}{pC_a\left(p^2 + \omega_{a2}^2\right)}, \qquad Z_b \triangleq \frac{p^2 + \omega_{b1}^2}{pC_b\left(p^2 + \omega_{b2}^2\right)}$$

$$Z_c \triangleq \frac{1}{pC_c} \qquad \text{and} \qquad Z_d \triangleq \frac{1}{pC_d}$$

It can be seen that the open and short circuit impedances of the lattice are of the form of a fourth degree (even) polynomial in p divided by fifth degree (odd) polynomial in p. Furthermore, the zeros of Z_{12} are given by

$$Z_a Z_d = Z_b Z_c$$

That is,

$$\frac{1}{p^2 C_a C_d}\left[\frac{p^2 + \omega_{a1}^2}{p^2 + \omega_{a2}^2}\right] = \frac{1}{p^2 C_b C_c}\left[\frac{p^2 + \omega_{b1}^2}{p^2 + \omega_{b2}^2}\right]$$

[3] It seems that these lattices are most useful for wideband filters.

and if we are not concerned with the single poles of loss at zero and infinite frequencies, this can be rearranged to

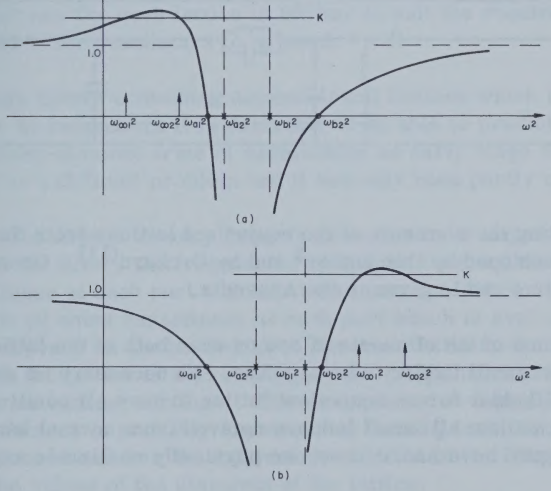
$$\frac{\left(\omega^2 - \omega_{a1}^{\;2}\right)\left(\omega^2 - \omega_{b2}^{\;2}\right)}{\left(\omega^2 - \omega_{a2}^{\;2}\right)\left(\omega^2 - \omega_{b1}^{\;2}\right)} = \frac{C_a\,C_d}{C_b\,C_c} \quad \triangleq K \tag{1}$$

This is a quadratic equation in ω^2 and with certain relationships between ω_{a1}, ω_{a2}, ω_{b1}, ω_{b2} and K it may have two positive roots $\omega_{\infty 1}^{\;2}$ and $\omega_{\infty 2}^{\;2}$ such that

$$\text{(a)}\ \omega_{\infty 1} < \omega_{\infty 2} < \omega_{a1} < \omega_{a2} < \omega_{b1} < \omega_{b2} \tag{2a}$$

$$\text{or}\quad \text{(b)}\ \omega_{a1} < \omega_{a2} < \omega_{b1} < \omega_{b2} < \omega_{\infty 1} < \omega_{\infty 2} \tag{2b}$$

Examples of these arrangements are shown in Figs. 2 (a) and (b).



(a)

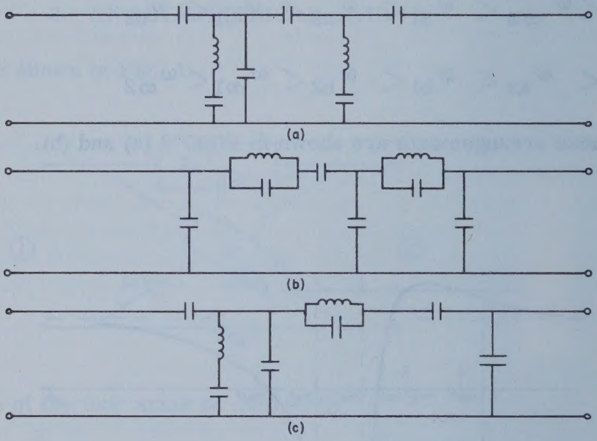
(b)

If, with either of these dispositions of the poles and zeros, $K < 1$, then a third arrangement of the positive roots $\omega_{\infty 1}^{\;2}$ and $\omega_{\infty 2}^{\;2}$ occurs —

$$\text{(c)}\ \omega_{\infty 1} < \omega_{a1} < \omega_{a2} < \omega_{b1} < \omega_{b2} < \omega_{\infty 2} \tag{2c}$$

This lattice is to be made equivalent to same sub two-port of a band-pass filter; in general, all the poles and zero of such a sub two-port will lie within or near to the pass band and therefore all the poles and zeros of the arms of an equivalent lattice will lie there also. Therefore, it can be seen that the arrangements of Equations 2(a), (b) and (c) may correspond to (a) two poles of loss in the lower stop band, (b) two poles of loss in the upper stop band or (c) one pole of loss in each stop band.

Thus the asymmetrical lattice shown in Figure 1 may be equivalent[4] to any of the three ladder circuits shown in Figure 3.



Methods of calculating the elements of the equivalent lattices from the elements of a ladder have been developed by this author[5] and by Orchard(11). Orchard's method is the more convenient and is given in the Appendix.

It is possible for some of the elements of one or even both of the lattices which are equivalent to a given small ladder to be negative. No necessary or sufficient conditions on the small ladder for an equivalent lattice to have all positive elements have been found but to date all small ladders derived from normal band-pass filters that have been designed have had at least one physically realizable asymmetrical lattice equivalent.

---

4. Since the lattice of Fig. 1 is asymmetrical, interchanging the numbering of the ports will result in different element values; therefore a given ladder network will be equivalent to two lattices — one as shown in Fig. 1 and one as shown but with the numbering of the ports reversed.
5. Internal memorandum of the G. P. O. Research Station, London, England. February 1964

3. BANDWIDTH CONSIDERATIONS

3.1 General

When designing conventional symmetrical lattices of eighth degree it is usual to arrange that these lattices have a triple pole of loss at zero or infinite frequency and remove the equal shunt or series inductors from inside the lattices to the ports. The three element arms of the lattices are then realized by resonators (plus, perhaps, capacitors).

It is well known that the bandwidths required of the resonators in such lattices are very much less than those of the overall filter. This is due to what may be loosely termed the broad-banding effect of the external inductors. For a conventional symmetrical lattice there are three possibilities -- two inductors (one in series and one in shunt) at each port, one inductor at each port (both in series or both in shunt) or no inductors.

External inductors have a broad-banding effect upon the asymmetrical lattices also but there is much more flexibility in the arrangement of the inductors; two, one or no inductors may be connected in series or shunt at one or both ports. The broad-banding effect is greatest for two inductors at each port, less for two at one port and one at the other, less still for one at each port, etc. A skilled designer will vary the arrangement for each lattice in a filter to suit the requirements of the overall filter and of the resonators to be used.

In order to design filters containing asymmetrical lattices which are to be realized wholly or partly by resonators it is desirable to be able to predict the capacity ratios[6] of the three-element arms of each lattice as early stage of the design as possible. This is a difficult problem and it has only been partly solved.

The capacity ratios depend on very many factors; the most important of these are the bandwidth of the filter, the frequencies of infinite loss and the amount of series or shunt capacitance at each port which is taken into the lattice. Further, the amount of series or shunt capacitance at each port which is available to be taken into the lattice depends on the bandwidth of the overall filter, the number of inductors by which the lattice is surrounded and the frequencies of infinite loss of any adjacent (ladder or lattice) sections. The comprehensive dependence of the capacity ratios on these factors is probably expressed by nothing less than all the steps of the synthesis process between the original description of the filter and the final calculation of the values of the elements of the lattice.

---

[6] From hereon it is more convenient to refer to the capacity ratio, r , of the three element arm than to its (relative) bandwidth; for a narrow band arm these are related by

$$r = \omega_R^2 / \left(\omega_A^2 - \omega_R^2\right) \simeq \omega_R / 2\left(\omega_A - \omega_R\right)$$

It is possible to greatly simplify this dependence in certain special cases if two rather crude assumptions are made -- firstly, that the values of the elements of an insertion loss designed filter are approximately equal to those of an image parameter designed filter with the same cut-off frequencies, poles of loss and configuration and secondly, that the capacitance ratios in a symmetrical lattice are approximately twice those in the equivalent (with a transformer) asymmetrical lattice. Then the capacity ratios in an asymmetrical lattice which is included in an insertion loss designed filter are approximately half those in the corresponding symmetrical lattice which is included in the image parameter filter.

This approach is only applicable to those sub two-ports for which there is a symmetrical image parameter two-port of the same configuration -- that is, those which have an inductor in series and shunt at both ports, in series at both ports, in shunt at both ports or no inductors. For the cases where there are dissimilar arrangements of inductors at the two-ports no method of predicting the capacity ratios of the asymmetrical lattice has been found.

In order to maximize the capacity ratios of the three-element arms of the lattice (or, put the other way, in order to maximize the bandwidth of a filter which can be designed with resonators with a given capacity ratio), it is best to include in the lattice all available capacity which is in series with or in shunt across a port. If however, two lattices are connected in tandem (with perhaps an inductor between), the partition of the central capacitor between the two lattices is variable but will considerably affect the capacity ratios in those two lattices. The partition which most nearly corresponds to that which would be achieved with the tandem connection of two image parameter sections will result in lattices with capacity ratios most nearly equal to those given below. However, there may be several reasons for deviating in practice from this particular partition and for a filter containing more than one lattice the expressions for the ratios should be used for determinations of feasibility only.

3.2 Lattices with one pole of loss in each stop band

If the approximation is made that the frequencies of infinite loss are geometrically symmetrical about the pass band (i.e., $\omega_{\infty 1}\,\omega_{\infty 2} = m_1\, m_2 = \omega_2$), then with $m = m_1$ or m_2, the capacity ratios of the arms of the symmetrical image parameter lattice, when divided by two, become........

with an inductor in series and shunt at each port,

$$r_a = r_b = \frac{\omega_2 \left(1+m\right)^2 \left(\omega_2 + m\right)^2}{2m\left(\omega_2+1\right)\left(\omega_2-1\right)^2\left(\omega_2+m^2\right)} \tag{2}$$

with an inductor in series at each port or an inductor in shunt at each port,

$$r_a = r_b = \frac{\omega_2 \left(1+m\right)^2 \left(\omega_2 + m\right)^2}{4m\left(\omega_2+1\right)\left(\omega_2-1\right)^2\left(\omega_2+m^2\right)} \tag{3}$$

---

* m_1 and m_2 are related to the poles of loss and the normalized upper cut-off frequency ω_2, by

$$m_i^2 = \frac{\omega_{\infty i}^2 - \omega_2^2}{\omega_{\infty i}^2 - 1}$$

and with no inductors

$$r_a = r_b = \frac{1}{2\,(\omega_2 - 1)} \tag{4}$$

3.3 Lattices with two poles of loss in the same stop band

In this case both m_1 and m_2 are < 1 or $> \omega_2$ and if the approximation is made that they are both equal to their geometric mean, m, then the capacity ratios of the arms of symmetrical image parameter lattice, when divided by two become.......

with an inductor in series and shunt at each port,

$$r_a = \frac{\omega_2 (1 + m)^2 (\omega_2 + m)^2}{2m^3 (\omega_2 + 1)(\omega_2 - 1)^2} \tag{5}$$

and $$r_b = \frac{(1 + m)^2 \; (\omega_2 + m)^2}{2m \; (\omega_2 + 1) \; (\omega_2 - 1)^2}\,, \tag{6}$$

with an inductor in series at each port

$$r_a = \frac{\omega_2^2 \, (1 + m)^2 \, (\omega_2^2 + 2m + m^2)}{4m^3 \, (\omega_2^2 - 1)^2} \tag{7}$$

and $$r_b = \frac{(1 + m)^2 \, (2m\omega_2^2 + \omega_2^2 + m^2)}{4m \; (\omega_2^2 - 1)^2}\,, \tag{8}$$

with an inductor in shunt at each port

$$r_a = \frac{\omega_2 (\omega_2 + m)^2 \; (2m\omega_2 + m^2 + 1)}{4m^3 \, (\omega_2^2 - 1)^2} \tag{9}$$

and $$r_b = \frac{(\omega_2 + m)^2 \, (m^2 \omega_2 + 2m + \omega_2)}{4m \; (\omega_2^2 - 1)^2} \tag{10}$$

and with no inductors

$$r_a = \frac{\omega_2^2 + m^2}{2m^2 \, (\omega_2^2 - 1)} \tag{11}$$

and $$r_b = \frac{1 + m^2}{2\left(\omega_2^{\ 2} - 1\right)} \tag{12}$$

It will be seen from equations (5) - (12) that if $m < 1$ then $r_a > r_b$ and if $m > \omega_2$ then $r_b > r_a$. The difference between r_a and r_b will increase with the sharpness (closeness to the pass band) of the poles of loss. If r_a and r_b are very different, only one of the three element arms of the asymmetrical lattice can be realized by a resonator and the other arm must be realized be an inductor and two capacitors. In this case, only the larger of each pair of r_a and r_b is of interest.

4. THE EFFECTS OF DISSIPATION AND INSTABILITY IN THE INDUCTORS

4.1 General

Dissipation and instability in the inductors may be treated in exactly the same way since the effect of both is to add in series with an inductor some extra impedance (resistive in the case of dissipation and reactive in the case of instability) the value of which is proportional to the inductance. That is, the expression pL for the impedance must be replaced by $pL\ (1 + \delta)$ where $\delta = -j/Q$ in the case of dissipation or is some real function of temperature or age in the case of instability.

The transformation (given in Appendix 1) from any of the ladder circuits shown in Fig. 3 to the asymmetrical lattice is derived on the basis of perfect inductors (i.e., $\delta = 0$) but is valid for any (constant) value of δ[7] — using, of course, the same δ for all four (two in the ladder and two in the lattice) inductors. Thus the effects of equal dissipation or instability in the inductors of any asymmetrical lattice are the same as the effects of the same dissipation or instability in the inductors of the equivalent ladder.

4.2 Ladders with one frequency of infinite loss in each stop band

In this case it is desirable to realize both three-element arms of the lattice by resonators and the improvement in the performance of the lattice will be the same as would be achieved in a ladder by using inductors of the same high quality as the resonators.

4.3 Lattices with two frequencies of infinite loss in one stop band

In this case it may be possible to realize only one of the three-element arms of the lattice by a resonator; the other must then be realized by a coil and two capacitors. Thus a sub two-port which produces two poles of loss close to the pass band is to be realized with one resonator and one coil; the results of such a realization must be examined most critically lest the dissipation and instability of the inductor largely nullify the good qualities of the resonator.

[7] In the equations which express the equivalences of the immittances of the two networks, the variable p^2 can be replaced by $p^2\ (1+\delta)$ without affecting the equalities.

Since the effects of equal δ are exactly the same in the ladder and the equivalent lattice, it is only necessary to consider the relative contributions to these overall effects of the δs of the two individual inductances of the lattice[8].

The only method of analyzing these relative sensitivities that has been found to be feasible and accurate is a conventional numerical analysis of the overall filter. Such an analysis has been performed on nearly all asymmetrical lattices with two poles of loss in the same stop band that have been designed. It has been found in all cases that the attenuation, A, of the overall filter at critical frequencies (the edge of the pass band nearest the poles of loss and the edge of the stop band) is very much less sensitive to dissipation and instability in the inductance (L_B) that must be realized by a coil than to dissipation and instability in the inductance (L_A) that may be realized as a resonator[9]. That is, at the critical frequencies

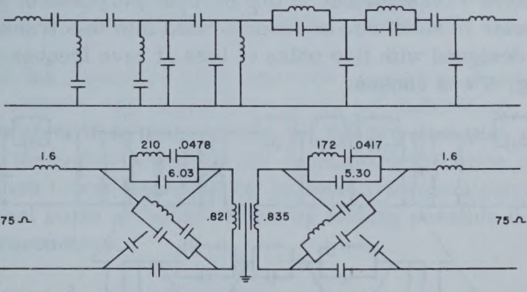
$$\frac{\delta A}{\delta L_B} \Big/ \frac{\delta A}{\delta L_A} \ll 1$$

This means that the improvement in stability and flatness of pass band that is achieved by realizing L_A as a resonator and L_B as a coil is almost as great as could be achieved by using two resonators (i.e., near perfect elements).

5. TWO EXAMPLES

To further investigate the bandwidths achievable with quartz crystals and to check the critical question of relative sensitivities, two filters were designed and analyzed.

The first was a simple filter with a pass band from 50 to 60 kHz and minima of 53 db attenuation below 48.8 kHz and above 61.3 kHz. A ladder realization and an equivalent network using two asymmetrical lattices are shown in Fig. 4.



50—60 KHZ B.P. (ELEMENT VALUES IN mH AND nF)

---

[8] It is well known (but not proven analytically) that in a ladder, imperfections in the inductor which produces the pole of loss closer to the pass band are more (but only slightly more) effective than imperfections in the other inductor.

[9] The singularities of the arm containing L_A are much closer to the critical edge of the pass band than are those of the arm containing L_B; therefore it is reasonable that variations of L_A should have more effect at this frequency.

The network was analyzed on a computer and then built at 75Ω impedance level using coils throughout[10].

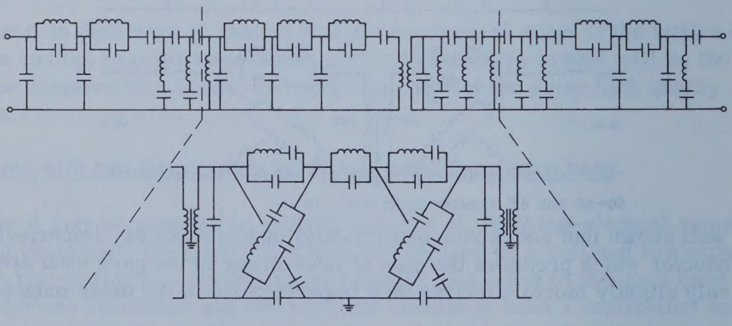
To investigate relative sensitivity the network was analyzed with dissipation in (a) only the broad-banding inductors, (b) only the wideband coils in the lattices and (c) only the narrow-band inductors in the lattices. On the breadboard the Q of the inductors of these three types was halved in turn and their relative contributions to the overall attenuation due to dissipation was calculated by extrapolation. The agreement between computed and measured results was very good and is summarized in Table 1.

TABLE 1

Relative Attenuation at edges of passband due to dissipation in —				
	Broad-banding inductors only	Inductors of wide-band arms of lattice only	Inductors of narrow-band arms only	All inductor
at 50 kHz	.16 db	.13 db	1.01 db	1.3 db
at 60 kHz	.14 db	.1 db	1.05 db	1.3 db

Thus at critical frequencies the change of response due to dissipation in the inductors in the wide-band arms of the lattices (to be realized by coils) is approximately one ninth that due to dissipation in the inductors in the narrow-band arms (to be realized by resonators). In other words, the effect of including the high-Q resonator in the lattice is approximately the same as would be achieved by multiplying the Q of both coils in the equivalent small ladder by ten.

The second example was a very difficult 60-108 kHz group interconnect filter. For a first attempt at using resonators in this filter the two poles of loss closest to the pass band on either side were realized by asymmetrical lattices. Preliminary calculations indicated that each lattice would need one broad-banding inductor. It is also useful for each lattice to have a transformer at one port for purposes of ground isolation; a broad-banding inductor in shunt can be incorporated into this transformer. Therefore the filter was designed with five poles of loss at zero frequency and the ladder configuration of Fig. 5 was chosen.



60-108 GROUP INTERCONNECT FILTER

[10] This postponed consideration of the problems associated with the high impedance of quartz crystals but was necessary for flexibility in the sensitivity investigation.

A small amount of shunt capacity was left at the outer ports of the two lattices to accommodate the anticipated distributed capacity of the high impedance secondary windings of the transformers. The central capacitor was divided so as to lead to approximately equal capacitance ratios in the narrow-band arms of the two lattices; these ratios were approximately 168:1 and 187:1 for the lattices realizing the two poles of loss in the upper and lower stop bands respectively.

The network was analyzed at the two edges of the pass band (60.166 and 107.75 kHz) with anticipated dissipation (Q = 550) only in the elements of the ladder network. Dissipation was introduced in turn into (a) the two broad-banding shunt inductors and the inductor of the parallel resonant circuit in the center[11], (b) the two inductors in the wide-band arms and (c) the two inductors in the narrow band arms. The results are summarized in Table 2.

TABLE 2

	Incremental attenuation at edges of pass band due to dissipation in —		
	Broad-banding inductors	Inductors of wide-band arms of lattices	Inductors of narrow-band arms of lattices
60.166 kHz	1.67 db	.57 db	2.91 db
107.75 kHz	2.45 db	.47 db	3.93 db

Thus at critical frequencies the change of response due to dissipation in the inductors in the wide-band arms is approximately one fifth (at 60.166) and one eighth (at 107.75) that due to dissipation in the inductors in the narrow band arms. That is, the effect of including high-Q coils in the lattices is approximately the same as would be achieved by multiplying the Q of both coils in the equivalent small ladder by six (at 60.166) and by nine (at 107.75).

The factor by which the equivalent Q is improved appears to depend on many parameters of the network; this dependence is not yet understood. This factor may be increased with further investigation but the two examples given and others analyzed by the author show that the asymmetrical lattice already offers between a five and ten to one improvement in sensitivity (to dissipation and instability) in critical ports of ladder filters by making possible the introduction of stable, high-Q resonators.

APPENDIX

The data for the required ladder to lattice transformation are the values of the eight elements of any one of the small ladders shown in Fig. 3. The frequencies of infinite loss are found and designated by $f_{\infty 1}$ and $f_{\infty 2}$.

---

[11] Because of the high impedance level of these inductors, the dissipation in them will be higher than it would be in the same inductors of the ladder. A Q of 200 was used.

The first step is to find and expand in partial fractions the open circuit admittance at port 1. The five elements of the Foster expansion of this admittance are defined by

$$\frac{1}{Z_{11}} = pC_3 + \frac{p}{L_1\left(p^2+\omega_1^2\right)} + \frac{p}{L_2\left(p^2+\omega_2^2\right)}$$

and $\omega_1^2 L_1 C_1 = \omega_2^2 L_2 C_2 = 1$

Three subsidiary parameters Δ, x and y are then defined by

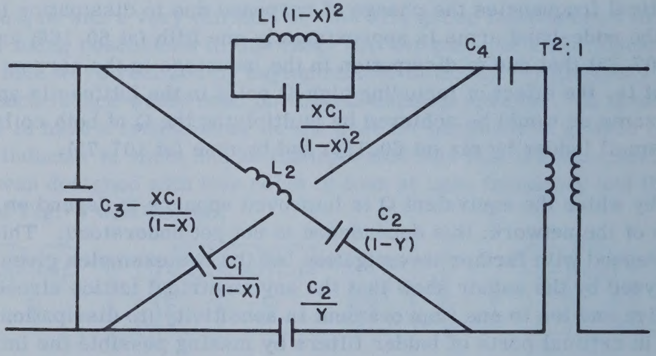
$$\Delta \triangleq \left(f_{\infty 1}^2 - f_1^2\right)\left(f_{\infty 2}^2 - f_1^2\right) + f_1^2\left(f_2^2 - f_1^2\right)$$

$$x \triangleq f_1^2\left(f_2^2 - f_1^2\right)/\Delta$$

$$y \triangleq f_1^2\left(f_{\infty 1}^2 - f_2^2\right)\left(f_{\infty 2}^2 - f_2^2\right)/\Delta f_2^2$$

The quantities (1-x) and (1-y) are also required and in order to conserve accuracy the subtractions should be performed algebraically.

The next step is to find the element values in the circuit of Fig. 6 so that it is a two-port equivalent of the given ladder.

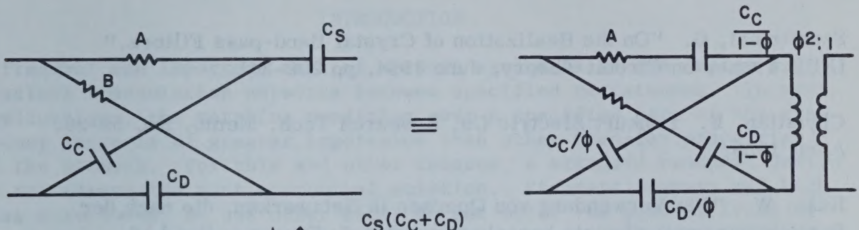


Seven of the element values are given by simple formulae as shown. The remaining capacitor, C_4, and the impedance ratio, T^2, of the ideal transformer are found by considering the short circuit impedance at port 2. Let this impedance behave like S_o/p as p tends to zero and like S_∞/p as p tends to infinity. Then

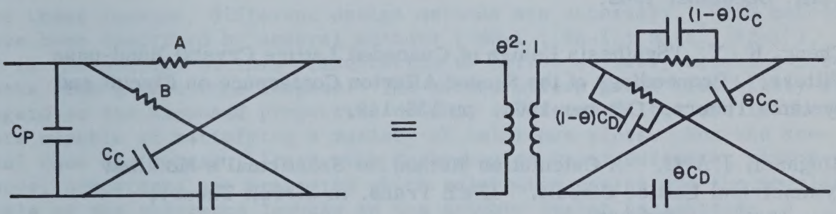
$$T^2\left(S_\infty - S_o\right) = (1-x)^2/C_1 + y^2/C_2$$

and $1/C_4 = T^2 S_o - y(1-y)/C_2$

The final step is to transform the circuit of Fig. 6 to the form of Fig. 1 by using the generalized Norton impedance transformations shown in Figs. 7 and 8.


A
B
CC
CD
CS
≡
A
CC/(1−φ)
φ2: 1
CC/φ
CD/(1−φ)
CD/φ

$$\phi \triangleq \frac{C_S(C_C+C_D)}{C_S(C_C+C_D)+C_C C_D}$$


A
B
CP
CC
≡
θ2: 1
(1−θ)CC
(1−θ)CD
θCC
θCD

$$\theta \triangleq \frac{C_C+C_D}{C_P+C_C+C_D}$$

REFERENCES

1. Poschenrieder, J. W. "Die Wellenparametertheorie als einfaches Hilfsmittel zur Realisierung von Quarzbandfiltern in Abzweigschaltung." NTZ. Band 12 (1959), No. 3.

2. Szentirmai, G. "On the Realization of Crystal Band-pass Filters." IEEE Trans. on Circuit Theory, June 1964, pp 299-301.

3. Christian, E. Lenkurt Electric Co. Research Tech. Memo. No. 59-20. April 1959.

4. Haas, W. "Die Verwendung von Quarzen in Netzwerken, die nach der Betriebsparametertheorie berechnet werden." Frequenz Band 16, May 1962, No. 5. pp 161-167.

5. Zorbalas, G. Thesis for degree of E.E., Stanford University, Calif. December 1962.

6. Cheng, K. Y. "Synthesis Design of Cascaded Lattice Crystal Band-pass Filters." Proceedings of the Second Allerton Conference on Circuit and Systems Theory. October 1964. pp 135-148.

7. Bingham, J.A.C. "A Calculation Method for Szentirmai's Modified Symmetrical Lattice Circuit." IEEE Trans. in Circuit Theory, June 1965. pp 284-285.

8. Lang, J. and Schmidt, C. E. "Crystal Filter Transformations." IEEE Trans. on Circuit Theory, September 1965. pp 454-457.

9. Christian, E. "Considerations for the Design of Crystal Filters." Proceedings of the Third Allerton Conference on Circuit and System Theory. October 1965. pp 806-816.

10. Bohse, J. "Die Verlustdampfung in Durchlassbereich von Wellenfiltern bei uteschiedlicher Spulengute." Frequenz, Band 12, December 1958, No. 12. pp 380-383.

11. Orchard, H. J. "A Design Method for an Asymmetrical Lattice Circuit." Lenkurt Electric Co., Advanced Development Tech. Memo. No. 63-19. November 1963.

BROAD-BAND MATCHING BY LOWPASS TRANSFORMATIONS

E. CHRISTIAN
ITT Telecommunications and N. C. State University, Raleigh, N. C.

E. EISENMANN
Telefunken A.G., Backnang, Germany

INTRODUCTION

A frequent and important task of network synthesis is the design of lossless transmission networks between specified resistances. In many applications, the matching condition over a specified part of the frequency range is of greater importance than other transfer properties of the network. For this and other reasons, a standard bandpass design is not always the most economical solution. Standard lowpass and highpass structures, on the other side, do not offer the possibilities of impedance transformations within the network except by the actual use of a transformer. It is therefore necessary to design such networks for specified terminations ([SA-1] ; [SA-2] ; [CE-1]). However, these design methods cause an excessive reflection coefficient at either terminal pair.

For these reasons, different design methods are necessary. Such methods have been described by several authors ([BO-1] ; [FA-1] ; [MA-1] ; [SZ-1]). It is the purpose of this presentation to derive suitably matching networks from a reference lowpass. The method offers great flexibility with regard to the stopband properties of this reference filter and is therefore capable of satisfying a variety of tolerance plots. For the special case of lowpass filters with Chebyshev or Cauer-Parameter performance, nomographs are presented which relate the degree and the modular angle of the reference lowpass to the primary design parameters:

(a) the boundaries of the matching range
(b) the upper boundary for the reflection coefficient in this range
(c) the ratio of the terminal resistances

As presented, the matching network will display a Chebyshev-type response over the matching range. However, a slight modification of the design method may be applied to yield a "maximally-flat" or any other type of performance over this range.

1. Statement of the Problem

Consider the operating condition of Fig. 1A in which a generator with an impedance R_1 shall be matched to a load impedance R_2 by means of a lossless matching network with lowpass structure. It is useful to normalize the entire circuit with respect to the generator impedance. (Fig. 1B). As long as no ideal transformer is part of the lowpass structure, at the frequency f = 0 the input impedances at the generator terminals will be equal to R_2 and r , respectively. The operating loss (i.e. the effective loss) at f = 0 is then:

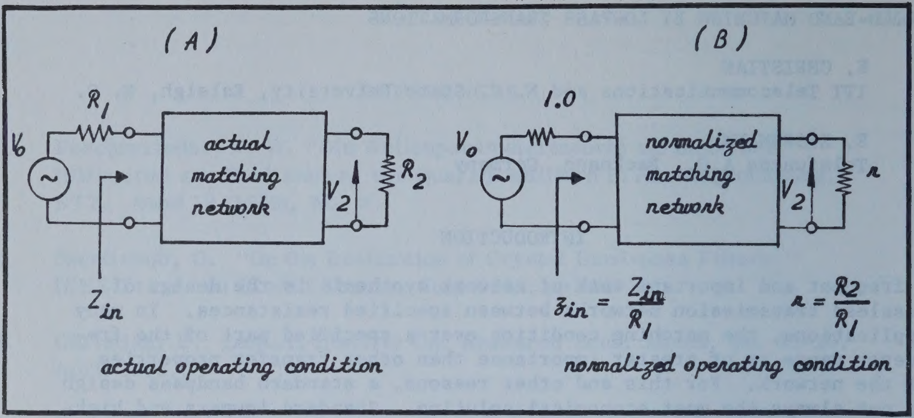

(A)
(B)
R_1
V_0
actual matching network
V_2
R_2
Z_{in}
actual operating condition
1.0
V_0
normalized matching network
V_2
r
$z_{in} = \frac{Z_{in}}{R_1}$
$r = \frac{R_2}{R_1}$
normalized operating condition

Fig. 1

$$A_r\,[db] = \begin{cases} = 10 \log_{10} \dfrac{P_{ref}}{P_2} = 10 \log_{10} \left[\tfrac{1}{2}\left(\sqrt{R_2/R_1} + \sqrt{R_1/R_2}\right)\right]^2 & (1) \\ = 10 \log_{10} \left[\tfrac{1}{2}\left(\sqrt{r} + 1/\sqrt{r}\right)\right]^2 & (2) \end{cases}$$

$$P_{ref} = \frac{|V_0|^2}{4R_1} \quad \text{maximally available power}$$

$$P_2 = \frac{|V_2|^2}{R_2} \quad \text{power dissipated in the load } R_2$$

Equ. (2) expresses the well known fact that the values A_r[db] are the same for reciprocal quantities of r. Therefore, one may restrict the following considerations for cases $r > 1$, i.e. $R_2 > R_1$, and reverse the resulting network in all other cases.

In order to achieve a specified degree of matching over a specified frequency range $f_{-p} \leq f \leq f_p$ it will be necessary to limit the operating loss A[db] by an upper boundary A_{max}[db] in this range. This boundary is related to an upper boundary ρ_{max} for the magnitude of the reflection coefficient according to the equation:

$$A_{max}\,[db] = 10 \log_1 \frac{1}{1 - \rho_{max}} \qquad (3)$$

$$\rho = \frac{R_1 - Z_{in}}{R_1 + Z_{in}} = \frac{1 - z_{in}}{1 + z_{in}} \quad \text{(reflection coeff.)} \qquad (4)$$

The condition at f = 0 , expressed by equations (1) or (2), and the upper boundary A_{max}[db] of equ. (3) postulate therefore the tolerance

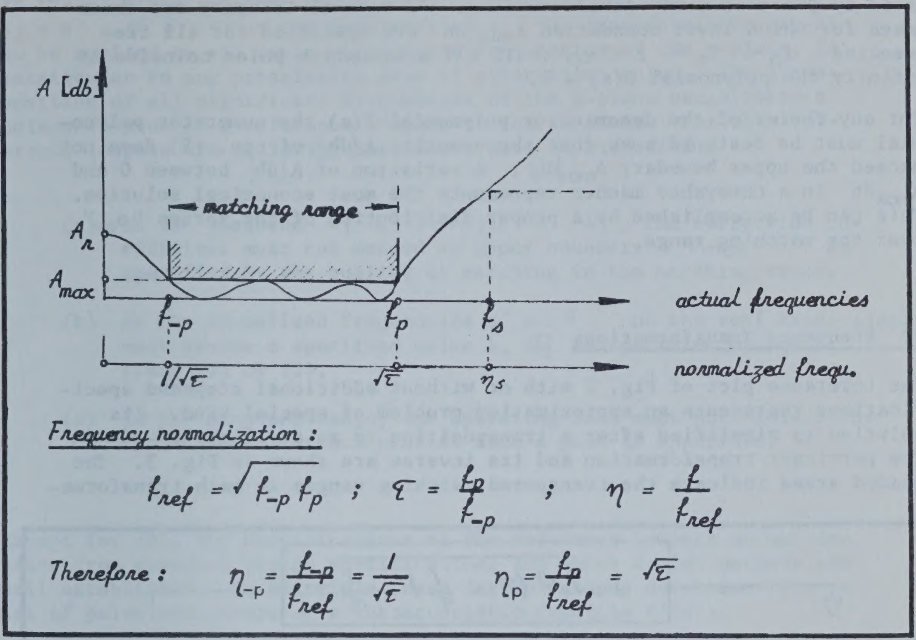

A [db]
matching range
A_r
A_{max}
actual frequencies
f_{-p}
f_p
f_s
$1/\sqrt{\tau}$
$\sqrt{\tau}$
η_s
normalized frequ.
Frequency normalization :
$f_{ref} = \sqrt{f_{-p} f_p}$; $\tau = \frac{f_p}{f_{-p}}$; $\eta = \frac{f}{f_{ref}}$
Therefore : $\eta_{-p} = \frac{f_{-p}}{f_{ref}} = \frac{1}{\sqrt{\tau}}$ $\eta_p = \frac{f_p}{f_{ref}} = \sqrt{\tau}$

Fig. 2

plot of Fig. 2 for the operating loss of the matching network. (If a highpass structure is preferred, equ. (2) must be satisfied at infinite frequency).

An amplitude performance A[db] satisfying the tolerance plot of Fig. 2 may be related to the following characteristic function K(s) of the matching network:

$$A\,[db] = 10 \log_{10}\left[1 + K(s)\,K(-s)\right]\Big|_{s=j\eta} \qquad (5)$$

$$K(s) = C\,\frac{\prod_{1}^{n}(s^2 + a_{0\nu}^2)}{P(s)} = C\,K_0(s) \qquad (6)$$

In a recent publication, ([AN-1]) a method was presented for the design of a suitable characteristic function K(s) by means of Zdunek's transformation. By this method, one arrives at a performance which satisfies equ. (2) at f = 0 and equ. (3) at f = f_p . The lower limit f_{-p} of the matching range, however, is not a design parameter and the resulting network may therefore exceed or violate the matching condition if both limits are specified.

The denominator polynomial P(s) of equ. (6) is necessarily even. It may be selected such that additional restraints may be satisfied in the stopband range $f > f_p$. Its degree must be equal to or smaller than (2n-2) if one wants to arrive potentially at a structure with no negative elements. The polynomial may be composed by root factors related to conjugate complex pairs of attenuation poles on the imaginary axis or to quadruplets in the complex plane. Proper locations may also be influenced by phase or dealy considerations. Of special interest are those cases for which lower boundaries $A_{min}[db]$ are specified for all frequencies $f_p < f_s \leq f < \infty$. If all attenuation poles coincide at infinity the polynomial $P(s) = 1$.

For any choice of the denominator polynomial $P(s)$ the numerator polynomial must be designed such that the quantity $A[db]$ of equ. (5) does not exceed the upper boundary $A_{max}[db]$. A variation of $A[db]$ between 0 and $A_{max}[db]$ in a Chebyshev manner represents the most economical solution. This can be accomplished by a proper distribution of the zeroes "a_{ov}" over the matching range.

2. Frequency Transformations

The tolerance plot of Fig. 2 with or without additional stopband specifications represents an approximation problem of special kind. Its solution is simplified after a transposition to an auxiliary s'-plane. The pertinent transformation and its inverse are shown in Fig. 3. The shaded areas indicate the transposed matching ranges in each transformation step. The overall transformation may be applied to all significant frequencies of the s-plane. Of special significance are the transpositions of the following normalized frequencies:

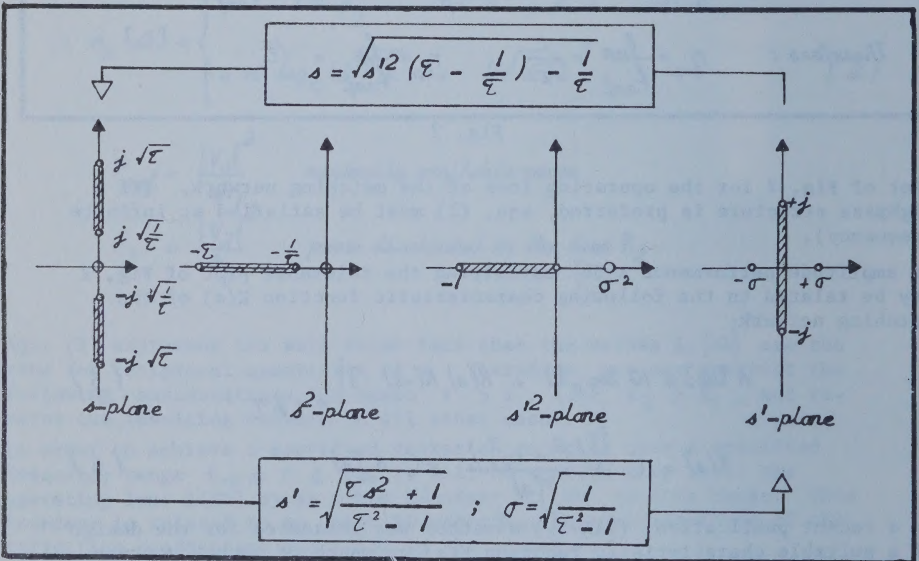

$s = \sqrt{s'^2\left(\tau - \frac{1}{\tau}\right) - \frac{1}{\tau}}$
$j\sqrt{\tau}$
$j\sqrt{\frac{1}{\tau}}$
$-j\sqrt{\frac{1}{\tau}}$
$-j\sqrt{\tau}$
$-\tau$
$-\frac{1}{\tau}$
-1
σ^2
$+j$
$-\sigma$
$+\sigma$
$-j$
s-plane
s^2-plane
s'^2-plane
s'-plane
$s' = \sqrt{\frac{\tau s^2 + 1}{\tau^2 - 1}}$; $\sigma = \sqrt{\frac{1}{\tau^2 - 1}}$

Fig. 3

$$\begin{array}{lcccc} s = & 0 & \pm j\frac{1}{\sqrt{\tau}} & \pm j\sqrt{\tau} & \infty \\ s' = & \pm\sigma & 0 & \pm j & \infty \end{array}$$

In the s'-plane the transposed matching range is the continuous range $-j \le s' = j\Omega \le +j$ on the imaginary axis. The transformation may also be applied to all other significant frequencies of the s-plane, in paraticular to any preselected sets of attenuation poles. The transposition of all significant frequencies of the s-plane establishes a tolerance plot of a reference lowpass in the s'-plane. For this reference lowpass the following restraints must be satisfied:

(a) In the passband $-j \le s' = j\Omega \le +j$, the reflection coefficient must not exceed an upper boundary ρ_{max} as specified by the quality of matching in the matching range.

(b) At the normalized frequencies $s' = \pm\sigma$ on the real axis, A[db] must assume a specified value A_r [db] as specified by equations (1) or (2).

(c) In the stopband range, the operating loss must not violate the transposed tolerance plot if such a plot is specified.

Except for (b), the specifications of the reference lowpass do not deviate from standard design specifications for which design methods are well established. In principle these design methods determine first a set of poles and zeroes of a characteristic function K'(s'):

$$K'(s') = C' K_0'(s') \qquad (8)$$

and then a suitable constant C'. For any selected poles and zeroes, this constant provides then one degree of freedom for an appropriate scaling of the loss performance. However, due to the condition (b) this degree of freedom must be used to make the operating loss equal to A_r [db] at $s' = \pm\sigma$. This postulate considerably reduces the possible choices of all those filters which could otherwise satisfy the specified tolerance plot.

As in the case of standard approximations, it is also possible to provide suitably graphs for the optimum selection of the degree and the modular angle if one restricts the considerations to lowpass filters of well defined performance, for instance to Cauer papameter or Chebyshev filters. The nomographs presented in the next subsection provide such an optimum selection based on the primary design parameters for matching, namely the ratio of the terminations and the matching range. (In these nomographs, the Chebyshev filters are treated as special cases of Cauer-parameter filters with $\Theta = 0^\circ$). Any filter selected according to the rules from these nomographs will satisfy condition (a) and (b). In a subsequent step, the selected type can then be investigated further if an eventual tolerance plot of the stopband is also satisfied. For this, well established graphs do exist ([CE-1]).

Due to the structure of the transformation formula, only even degree characteristic functions are permitted for the reference lowpass. This postulates, eventually, only antimertical matching networks.

3. Design Nomographs for Cauer-Parameter and Chebyshev Performance.

For the particular case of a reference lowpass with Cauer-parameter or Chebyshev characteristics, the function K (s') of equ.(8) is of the following structure:

$$K_0'(s') \begin{cases} = \dfrac{\prod_1^n (s'^2 + a'^2_{0v})}{\prod_1^{n-1} (s'^2 + a'^2_{\infty v})} & \text{for Cauer-parameter filters} \\ = \prod_1^n (s'^2 + b'^2_{0v}) & \text{for Chebyshev filters} \end{cases} \qquad (9)$$

The poles and zeroes of these functions may easily be calculated by well established formulas. The respective attenuations A_{00}[db] at s' = 0 and $A_{0\sigma}$[db] at s' = σ related to $K_0'(s')$ are:

$$A_{00}[db] = 10 \log_{10} [1 + |K_0'(0)|^2] \qquad (10)$$

$$A_{0\sigma}[db] = 10 \log_{10} [1 + |K_0'(\sigma)|^2] \qquad (11)$$

These quantities are functions of the degree and the modular angle only.

Augmenting $K_0'(s')$ by a constant C' yields the actual characteristic function K'(s') of the reference lowpass. For this reference lowpass it is postulated that the operating loss at s' = 0 and at s' = $\pm\sigma$ assumes the specified values A_{max}[db] and A_r [db], respectively:

$$A_{max}[db] = 10 \log_{10} [1 + C'^2 |K_0'(0)|^2] \qquad (13)$$

$$A_r\,[db] = 10 \log_{10} [1 + C'^2 |K_0'(\sigma)|^2] \qquad (14)$$

Equations (10) and (13), or (11) and (14) yield then:

$$C' = \sqrt{\frac{10^{0.1 A_{max}} - 1}{10^{0.1 A_{00}} - 1}} = \sqrt{\frac{10^{0.1 A_r} - 1}{10^{0.1 A_{0\sigma}} - 1}} \qquad (15)$$

Therefore:

$$\sqrt{\frac{10^{0.1 A_{max}} - 1}{10^{0.1 A_r} - 1}} = \sqrt{\frac{10^{0.1 A_{00}} - 1}{10^{0.1 A_{0\sigma}} - 1}} \qquad (16)$$

The left part of equ.(16) is a function of the primary design parameters alone and is independent of the filter type, its degree and modular angle. The opposite is true for the right hand side of this equation. This is only possible if either side of the equation is equal an auxiliary constant L^2:

$$\sqrt{\frac{10^{0.1 A_{max}} - 1}{10^{0.1 A_r} - 1}} = L^2 = \sqrt{\frac{10^{0.1 A_{00}} - 1}{10^{0.1 A_{0\sigma}} - 1}} \qquad (17)$$

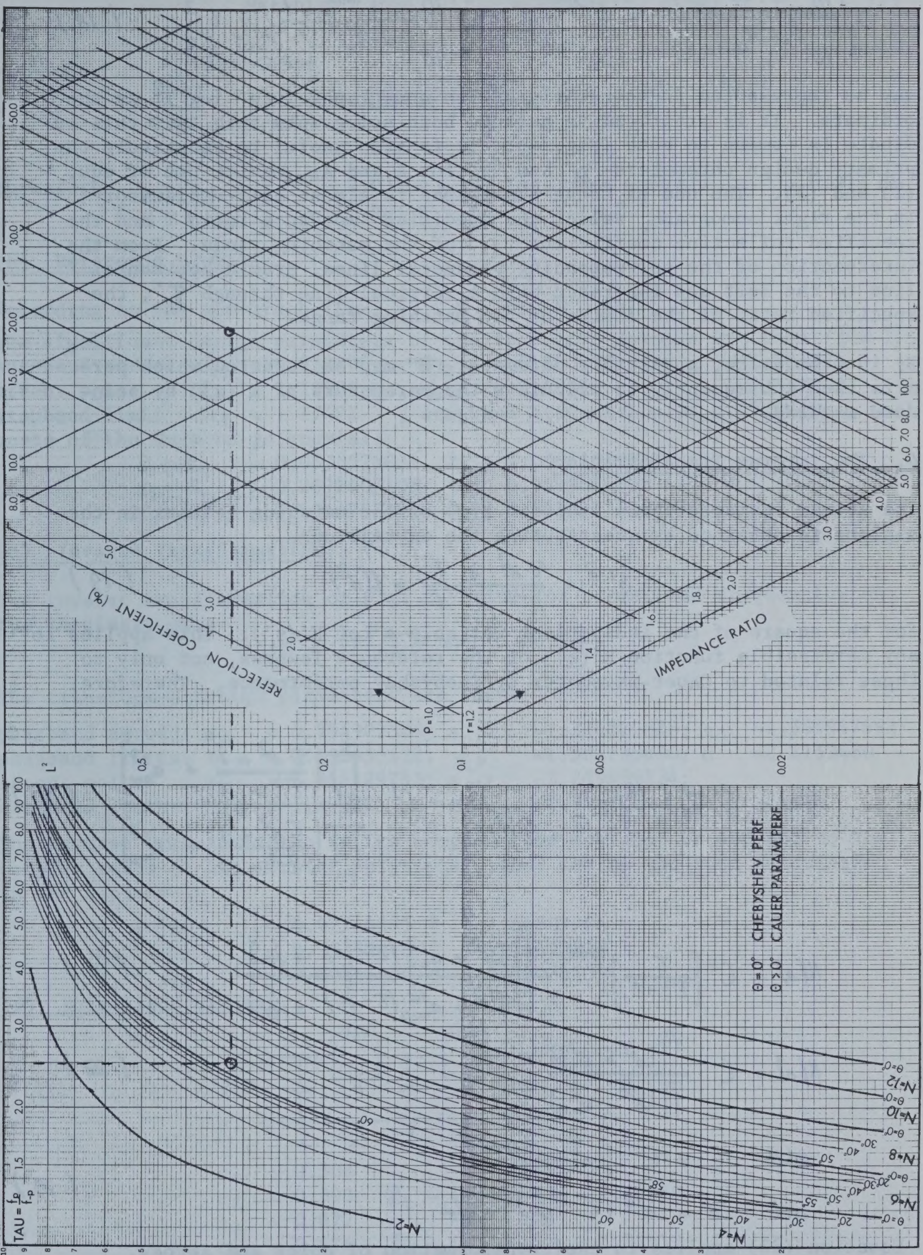

TAU = $\frac{f_p}{f_{-p}}$
L^2
REFLECTION COEFFICIENT (%)
IMPEDANCE RATIO
ρ=1.0
r=1.2
θ = 0° CHEBYSHEV PERF
θ > 0° CAUER PARAM PERF
N=2
N=4
N=6
N=8
N=10
N=12
θ=0°
10°
20°
30°
40°
50°
55°
58°
60°
1.4
1.6
1.8
2.0
3.0
4.0
5.0
6.0
7.0
8.0
10.0
1.5
15.0
20.0
30.0
50.0
0.5
0.2
0.1
0.05
0.02

Fig. 4

The two equations:

$$\sqrt{\frac{10^{0.1\,A_{oo}}-1}{10^{0.1\,A_{o\sigma}}-1}} = L^2 \quad \text{and} \quad \sqrt{\frac{10^{0.1\,A_{max}}-1}{10^{0.1\,A_r}-1}} = L^2 \qquad (18)$$

may be represented independently by two nomographs with log L^2 as the ordinate. It is practical to arrange both nomographs such that this ordinate is common to both (Fig. 4). It is also useful to substitute the primary design parameters "ρ" and "r" for A_r and A_{max} making use of equ. (2) and (3). The use of these nomographs is almost self-explanatory: In the right nomograph, select the specified impedance ratio r and the specified reflection coefficient "ρ" of the matching range. From the intersection of the corresponding parameter lines, draw a horizontal line to the left and intersect with a vertical line through "τ", the specified ratio of the limits of the matching range.

By interpolation, a suitable parameter "Θ" may be selected which permits then subsequently to calculate the parameters of K'(s') by standard methods.

4. The Inverse Transformation.

The actual characteristic function K(s) of the matching network in the s-plane is obtained by substituting the expression:

$$s'^2 = \frac{\tau s^2 + 1}{\tau^2 - 1} = f(s^2) \qquad (19)$$

for the square of the independent variable s' of K'(s'). This applies to any permissible function K'(s') of the reference lowpass, not only to those related to Cauer-parameter or Chebyshev performance. Therefore:

$$K(s) = C' K_o'(s'^2)\Big|_{s'^2 = f(s^2)} = C' \frac{\prod_{\nu=1}^{n}\left[\frac{\tau s^2+1}{\tau^2-1} + a'^2_{o\nu}\right]}{\prod_{\mu=1}^{m}\left[\frac{\tau s^2+1}{\tau^2-1} + a'^2_{\infty\mu}\right]} \qquad (20)$$

$$K(s) = C'\left[\frac{\tau}{\tau^2-1}\right]^{n-m} \frac{\prod^{n}\left[s^2 + a'^2_{o\nu}\left(\tau-\frac{1}{\tau}\right) + \frac{1}{\tau}\right]}{\prod^{m}\left[s^2 + a'^2_{\infty\mu}\left(\tau-\frac{1}{\tau}\right) + \frac{1}{\tau}\right]} \qquad (21)$$

$$K(s) = C\,\frac{\prod^{n}\left[s^2 + a^2_{o\nu}\right]}{\prod^{m}\left[s^2 + a^2_{\infty\mu}\right]} \qquad (22)$$

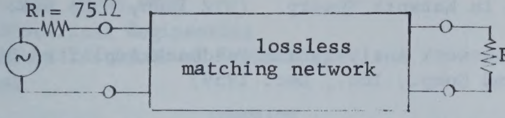
$$\text{with} \quad C = C'\left[\frac{\tau}{\tau^2-1}\right]^{n-m}$$

$$\text{and} \quad a^2 = a'^2\left(\tau-\frac{1}{\tau}\right) + \frac{1}{\tau} \qquad (23)$$

Equation (23) applies to the transformation of poles and zeroes.

5. Numerical Example.

(a) Specifications : Terminal imped. : $R_1 = 75\,\Omega$ $R_2 = 150\,\Omega$
Matching range : $f_{-p} = 200$ kc; $f_p = 500$ kc; $\rho_{max} = 12\%$



therefore: $r = R_2 / R_1 = 2.0$; $A_r = 10 \log [\tfrac{1}{2}(\sqrt{r} + 1/\sqrt{r})]^2 = 0.511$ db
$f_{ref} = \sqrt{f_{-p}/f_p} = 316$ kc; $\tau = f_p/f_{-p} = 2.5$

(b) Reference Lowpass:

According to the nomograph (Fig. 4): $N = 6$; $\theta = 60^\circ$

The significant design parameters of the reference lowpass are then ([CE-1])

passband zeroes: $\grave{a}_{01} = 0.332292$; $\grave{a}_{02} = 0.803358$; $\grave{a}_{03} = 0.981774$
atten. poles : $\grave{a}_{\infty 1} = 1.19686$; $\grave{a}_{\infty 2} = 1.48983$; $C' = 4.6522317$
stopband limit $\Omega_s = 1.17422$; stopband attenuation: $A_{min} = 32.0$ db
roots of the Hurwitz polynomial:
$\grave{a}_1 \pm j\grave{b}_1 = 0.0578579671 \pm j\,1.0504403983$
$\grave{a}_2 \pm j\grave{b}_2 = 0.2672409105 \pm j\,0.9499866833$
$\grave{a}_3 \pm j\grave{b}_3 = 0.6514237489 \pm j\,0.4614112719$

(c) Inverse Transformation:

The inverse transformation of Fig. 3 yields for normalized frequencies on the imaginary axis:

$$\eta^2 = \Omega^2(\tau - 1/\tau) + 1/\tau$$

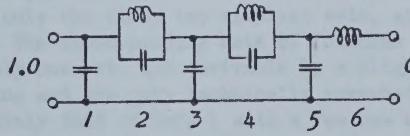
The transformed design parameters are then:

passband zeroes: $a_{01} = 0.794907380$; $a_{02} = 1.324879828$; $a_{03} = 1.556967691$
passband limits: $\eta_{-p} = 0.632455532$; $\eta_p = 1.581138830$; $C = 2.2153480$
atten. poles : $a_{\infty 1} = 1.846129763$; $a_{\infty 2} = 2.249699134$;
stopband limit : $\eta_s = 1.17422$ $\eta_s/\eta_p = 1.148122$

The inverse transformation in its complex form may also be applied to the complex roots of the Hurwitz polynomial, thus saving the root calculation in the s-plane:

$a_1 \pm jb_1 = 0.07744193 \pm j\,1.64807766$
$a_2 + jb_2 = 0.35382416 \pm j\,1.50678838$
$a_3 + jb_3 = 0.80846475 \pm j\,0.78074639$

(d) Normalized Circuit



$c_1 = 0.508$
$c_2 = 0.393$ $\ell_2 = 0.502$ $\eta_2 = 2.24969$
$c_3 = 0.877$
$c_4 = 0.924$ $\ell_4 = 0.317$ $\eta_4 = 1.84612$
$c_5 = 1.06$
$\ell_6 = 0.407$

The network provides a step-down of the impedance. It should therefore be reversed and denormalized with $R_{ref} = 150$ Ohms and $f_{ref} = 316$ kc.

BIBLIOGRAPHY

[AN-1] Antreich, K. and Gleissner, E.: On Some Applications of Zdunck's Transformation in Network Theory. (NTZ 1966, H.2, pgs. 101-105)

[BO-1] Bode, H. W.: Network Analysis and Feedback Amplifier Design. (D. van Nostrand Comp., Inc., Dec. 1959)

[CE-1] Christian, E. and Eisenmann, E.: Filter Design Tables and Graphs. (To be Published by J. Wiley & Sons, Inc.)

[FA-1] Fano, R. M.: Theoretical Limits on Broad Band Matching of Arbitrary Impedances. (J.F.I., Jan. 1950, pgs. 58-83, Feb. pgs. 139-154)

[MA-1] Matthei, G. L.: Tables of Chebyshev Impedance Transforming Networks of Lowpass Filter Transform.)Proc. IEEE, Aug. 1964, pgs. 939-963)

[SA-1] Saal, R. and Ulbrich, A. E.: On the Design of Filters by Synthesis. (IRE Trans. CT, Dec. 1958).

[SA-2] Saal, R.: On the Design of Lowpass and Highpass Filters with Unequal Terminations. (NTZ Jhrg. 17, H. 11, pgs. 555-59)

[SZ-1] Szentirmai, G.: Bandpass Matching Filter in the Form of Polynomial Lowpass Filter. (IEEE Trans. CT, 11, No. 1, March 1964)

EXPLICIT FORMULAS FOR BUTTERWORTH AND CHEBYCHEFF LADDER FILTERS WITH NEW PATTERNS OF REFLECTION ZEROS*

I.NAVOT, B.GONEN AND Z.PERI
Faculty of Electrical Engineering
Technion, Israel Institute of Technology
Haifa, Israel

ABSTRACT

An outstanding problem in network theory for the last forty years has been that of deriving explicit fomulas for the element values of Butterworth and Chebycheff ladder filters. Earlier attempts at this problem have been superseded in 1951 by Takahasi's unified solution to a sub-class of filter problems[1]. However, even Takahasi's solution was far from solving the problem in its full generality since, as is well known from Darlington's theory, a Butterworth or Chebycheff filter of order n possesses 2^m, $m=[(n-1)/2]$, essentially different ladder realizations depending on the pattern of reflection zeros chosen. Takahasi's solution pertains to the pattern where all these zeros are restricted to the left or right half complex frequency plane or to Weinberg's closely related "alternating pattern" (which exists only for n odd). Extending Takahasi's ideas we firstly classify, with the aid of the Routh-Hurwitz Theorem[2], all admissible patterns according to their degree of complexity. This immediately reveals that besides Weinberg's and Takahasi's patterns, which are the simplest, other patterns, only slightly more complicated, exist for which the required formulas can readily be derived by elementary techniques. Turning then to more complicated patterns we develop an algorithm, whose application becomes more laborious with the increase in the degree of complexity of the corresponding pattern, which yields the required formulas in these cases also. The catalogueing of formulas pertinent to all admissible patterns for arbitrary large n is, however, not yet complete.

1. INTRODUCTION

The problem we consider is presented as a mathematical one, that of deriving explicit formulas for the coefficients in a certain continued fraction expansion of a given rational function in the complex frequency variable s. The existence of this particular expansion and its interpretation as a driving point immitance of a reactive ladder terminated in a dissipative element are well known from network theory. Lack of space permits us only to outline the basic steps in the derivation of the sets of formulas pertinent to Butterworth filters with the first three simplest patterns of reflection zeros (after those of Weinberg and Takahasi). We list only the first two simplest sets, along with Takahasi's known one for comparison. The corresponding sets of formulas for Chebycheff's filters, of which we list only one set, are derivable by a slight modification of the same basic reasoning and are only technically somewhat more complicated. Our notation follows mainly that of Ref. 1 with a few new additions.

---

*The contributions of the second and third named authors are based on theses, supervised by the first named author, to be submitted to the Senate of the Israel Institute of Technology, Haifa, Israel, in partial fulfilment of the requirements for the M.Sc. degree.

Formulation of problem

Our point of departure is the function

$$\rho(s,x)=\frac{H(s,x)}{G(s)}\,, \qquad -1<x<1 \tag{1}$$

where $G(s)$ and $H(s,x)$ are monic real polynomials in s of degree n; $n=1,2,\ldots$, defined by

$$G(s)=\prod_{v=1}^{n}(s-j\varepsilon^{2v-1})\,, \qquad \varepsilon=\exp.(j\pi/2n) \tag{2}$$

and

$$H(s,x)H(-s,x)=(-)^n s^{2n}+x^{2n}. \tag{3}$$

The s-plane zeros of $H(s,x)$, referred to as the "reflection zeros", are located on a circle of radius $|x|$ concentric with the unit circle which is the locus of the zeros of $G(s)$. For $x\neq 0$ there are 2^m, $m=[(n+1)/2]$, different ways of choosing the zeros of $H(s,x)$. This number is reduced to half its value if we consider two patterns of reflection zeros derived from each other on replacing x by $-x$ as not essentially different. For the degenerate case $x=0$ the various patterns become undistinguishable. We term the patterns "singlets" (i), "doublets" (i,j) etc., according as a single pair, two pairs etc., of complex conjugate reflection zeros are in one half plane (right or left) while the remaining ones are in the other half plane (left or right respectively). The indices in brackets, (i), (i,j) etc., are identification tags indicating which pair, or pairs, of zeros on the reflection zeros' circle, counted from the imaginary axis, are the ones in case. Thus, for Takahasi's pattern and for the singlets (1), (2) and for the doublet (1,2) the forms assumed by $H(s,x)$, $-1<x<1$, are

$$\left.\begin{aligned}
&H(s,x)=\prod_{v=1}^{n}(s-xj\varepsilon^{2v-1})\,, \quad n=1,2,\ldots\\
&H(s,x)=(s-xj\varepsilon^{-1})(s+xj\varepsilon)\prod_{v=2}^{n-1}(s-xj\varepsilon^{2v-1})\,, \quad n=3,4,\ldots\\
&H(s,x)=(s-xj\varepsilon^{-3})(s+xj\varepsilon^{3})\prod_{v=1}^{n}(s-xj\varepsilon^{2v-1})\,, \quad v\neq 2,n-1\,, \quad n=4,5,\ldots\\
&H(s,x)=(s-xj\varepsilon^{-3})(s-xj\varepsilon^{-1})(s+xj\varepsilon^{3})(s+xj\varepsilon)\prod_{v=3}^{n-2}(s-xj\varepsilon^{2v-1})\,, \quad n=5,6,\ldots
\end{aligned}\right\} \tag{4}$$

respectively. The continued fraction expansion of interest is that of the positive real function

$$W(s,x)=\frac{1+\rho(s,x)}{1-\rho(s,x)}=\frac{G(s)+H(s,x)}{G(s)-H(s,x)}\quad,\qquad -1<x<1 \tag{5}$$

and its form corresponds to a driving point immitance of a reactive ladder with a pole at infinity and terminated in a dissipative element. (Recall that for $n>1$ the even part of $W(s,x)$ as a function of s vanishes at infinity only). Associated with this expansion is the Sturm sequence of polynomials in s (with reversed signs) $P_o(s,x)$, $P_1(s,x),\ldots,P_n(s,x)$ given by

$$\begin{aligned}
&P_o(s,x)=G(s)+H(s,x) && =a_{oo}(x)s^n+\ldots+a_{on}(x)\\
&P_1(s,x)=G(s)-H(s,x) && =a_{11}(x)s^{n-1}+\ldots+a_{1n}(x)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots && \\
&P_m(s,x)=P_{m-2}(s,x)-a'_{m-1}(x)sP_{m-1}(s,x) && =a_{mm}(x)s^{n-m}+\ldots+a_{mn}(x)\\
&\qquad m=2,3,\ldots,n\ , \qquad a'_{m-1}(x)=\frac{a_{m-2,m-2}(x)}{a_{m-1,m-1}(x)} &&
\end{aligned} \tag{6}$$

with $a_{vv}(x)>0$; $v=0,1,\ldots,n$, for $-1<x<1$. Evidently, the derivation of explicit formulas for the Sturm coefficients $a_{vv}(x)$ and $a_{n-1,n}(x)$ is sufficient for the solution of our problem.

Two conclusions from (6), which will be capitalized upon in the sequel, may now be stated:

(1) The Sturm sequence of polynomials associated with the continued fraction (with unit partial numerators) of the positive real odd function

$$W_1(s,x)=\frac{\frac{1}{2}[P_o(s,x)+(-)^nP_o(-s,x)]}{\frac{1}{2}[P_1(s,x)-(-)^nP_1(-s,x)]} \tag{7}$$

is obtained from (6) on deleting from each row of the triangular array of coefficients on its right hand side the second, fourth etc. coefficient, namely, $a_{v,v+1+2i}$, $v=0,1,\ldots,n-1$, $i=0,1,\ldots,[(n-v-1)/2]$. (Note that by their law of formation the remaining coefficients are algebraically independent of the deleted ones).

(2) If for a certain value of x, say x_u, real or complex and not in $-1<x<1$, $G(s)$ and $H(s,x_u)$ have f common zeros $s_{u1}, s_{u2}, \dots, s_{uf}$, $f=1,2,\dots,n$, and $a_{vv}(x_u) \neq 0$ for $v=0,1,\dots,n-f$ then $s_{u1}, s_{u2}, \dots, s_{uf}$ are also zeros of $P_o(s,x_u)$, $P_1(s,x_u), \dots, P_{n-f}(s,x_u)$ whereas $P_{n-f+1}(s,x_u)$ vanishes identically in s, i.e. $a_{n-f+1,n-f+1+i}(x_u)=0$ for $i=0,1,\dots,f-1$. The values of x_u of interest in this connection are, obviously, $x_u=\varepsilon^{2u}$, $u=-n+1,-n+2,\dots,-1,0,1,\dots,n-1,n$. The function f(u), which is an even one, is a characteristic of the pattern and no two essentially different patterns have the same f(u). It plays a fundamental role in the derivation of the required explicit formulas.

2. THE ROUTH-HURWITZ DETERMINANTS FOR $a_{v,v+2i}$

Closed form expressions for the coefficients a_{vv}, $a_{v,v+2}, \dots$, $v=2,3,\dots,n$ in terms of a_{oo}, $a_{o2}, \dots$ and a_{11}, $a_{13}, \dots$ are known [2] to exist as quotients of sub-determinants formed from the Hurwitz square matrix of order n

$$M = \left\| \begin{matrix} a_{11} & a_{13} & a_{15} & \cdots\cdots & a_{1,2n-1} \\ a_{oo} & a_{o2} & a_{o4} & \cdots\cdots & a_{o,2n-2} \\ 0 & a_{11} & a_{13} & \cdots\cdots & a_{1,2n-3} \\ 0 & a_{oo} & a_{o2} & \cdots\cdots & a_{o,2n-4} \\ 0 & 0 & a_{11} & \cdots\cdots & a_{1,2n-5} \\ \cdot & \cdot & \cdot & \cdots\cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots\cdots & \cdot \end{matrix} \right\| \qquad \begin{matrix} (a_{1i} \text{ and } a_{oi}=0 \\ \text{for } i>n) \end{matrix} \qquad (8)$$

associated with the polynomial in s

$$\tfrac{1}{2}[P_o(s,x)+(-)^n P_o(-s,x)] + \tfrac{1}{2}[P_1(s,x)-(-)^n P_1(-s,x)].$$

(For simplicity's sake we do not exhibit the polynomial dependence on x of the elements of M, except $a_{oo}=2$, which is the case for most patterns. An exception to this occurs in the case of Weinberg's "alternating pattern" where the only element of M dependent on x is a_{1n}). The pertinent expressions are

$$a_{v,v+2i}=\frac{D_{v,v+2i}}{D_{v-1,v-1}} \quad , \quad v=1,2,\dots,n \quad , \quad i=0,1,\dots,[(n-v)/2] \qquad (9)$$

where $D_{v-1,v-1}$ is the principal sub-determinant formed from the first v-l rows of M whereas $D_{v,v+2i}$ is the sub-determinant obtained from D_{vv} on replacing its last column by the first v elements of the (v+i)th column of M.

3. PATTERN CLASSIFICATION AND THE DERIVATION OF THE SIMLEST FORMULAS

We are now in a position to propose a practical way of classifying the various patterns according to their degree of complexity and to describe a simple algorithm for deriving the explicit formulas pertinent to the simplest patterns. For patterns with a higher degree of complexity a more complicated algorithm has been developed which will be described elsewhere. The catalogueing of formulas pertinent to all possible patterns for arbitrary large n is, however, not yet complete.

Firstly we obtain an upper bound , attained by most patterns, for deg.$D_{vv}(x)$ (the degree of D_{vv} as a polynomial in x). To this end we assign to each non vanishing element a_{iv}, i=0,1, of M in (8) a weight, $w(a_{iv})$, equal to its second subscript; to each non vanishing term in the expansion of D_{vv} we assign a weight equal to the sum of the weights of its factors, e.g. $w(a_{11}a_{02}a_{13})=6$, and to D_{vv} we assign a weight equal to the maximal weight of its non vanishing terms. A little consideration shows (either straightforwardly or by mathematical induction) that all the non vanishing terms of D_{vv} are isobaric. Since deg.$a_{iv}(x)\leqslant w(a_{iv})$, i=0,1, we have

$$\text{deg.}D_{vv}(x)\leqslant w(D_{vv})=\frac{v(v+1)}{2} \quad , \quad v=1,2,\dots,n. \qquad (10)$$

We then factorize $D_{vv}(x)$ as far as possible, using conclusion (2) at the end of section 1, extracting from it groups of divisors of the polynomials $1-x^n$ and $1+x^n$ to a multiplicity depending on the particular pattern at hand. (The factorization can be viewed in retrospect from next section, recalling that $D_{vv}(x)=a_{11}(x)a_{22}(x)\dots a_{vv}(x)$). What remains of $D_{vv}(x)$, denoted by $y_{vv}(x)$ and termed the "kernel polynomial of $D_{vv}(x)$", is generally prime to $1-x^{2n}$ except in degenerate cases, i.e. for certain values of n at which the pattern considered actually belongs to two pattern classes, e.g. singlets (1) and (2) for n=4, singlet (3) and doublet (1,2) for n=6, etc. The maximal degree of $y_{vv}(x)$, obtained from (10) on subtracting from deg.$D_{vv}(x)$ the sum of the degrees of the extracted divisors of $1-x^{2n}$, determines the complexity of the pattern. Thus, for Takahasi's pattern and for the singlets (1), (2), (3) and the doublet (1,2) it turns out, from an examination of the function f(u) introduced in conclusion (2), that the degree of the kernel polynomial $y_{vv}(x)$ of $D_{vv}(x)$, v=1,2,...,n, does not exceed 0,1,3,5 and 4 respectively.

Turning now to the derivation of the required explicit formulas, it follows from the preceding that the missing link is the determination of $y_{vv}(x)$. Appealing to Lagrange's interpolation formula, it is evident that knowledge of $y_{vv}(x)$ at a number of distinct non zero values of x equal to its maximal degree completes the solution of our problem since $y_{vv}(0)$, which is the same for all patterns, is known from the "equal resistance terminations" case. For deg.$y_{vv}(x)$ $\leqslant 4$ this can easily be accomplished with the aid of "Orlando's formula"[2] which yields $D_{n-f-1,n-f-1}(x_u)$ and $D_{n-f,n-f}(x_u)$, $n-f-1 \geqslant 1$, in terms of the zeros $s_{u,f+1}$, $s_{u,f+2}$, ..., s_{un} of G(s) at which $H(s,x_u)$ does not vanish. Replacing the value of u, say u_1, corresponding to a certain definite value of n-f-1 by another value of u, say u_2, such that $n-f(u_2)=n-f(u_1)-1$, Orlando's formula yields $D_{n-f-1,n-f-1}(x)$ at $x=\xi^{\pm 2u_1}$ and $x=\xi^{\pm 2u_2}$. $y_{n-f-1,n-f-1}(x)$ at these values of x is then obtained from that of $D_{n-f-1,n-f-1}(x)$ on dividing the latter by a readily evaluated non vanishing expression. Finally, the formulas for $y_{vv}(x)$ v=1,2,...,n become considerably simpler if we normalize them so that $y_{vv}(0)=1$, i.e. we also extract from $D_{vv}(x)$ the numerical factor $D_{vv}(0)=(2S_1)(2S_2)...(2S_v)$ where S_v is defined in next section.

It is interesting to remark that the problem of synthesizing all resistance terminated reactive ladders corresponding to a prescribed transfer function with all transmission zeros at infinity (of which Butterworth and Chebycheff low-pass ladder filters are special cases) is equivalent to synthesizing a certain sub-class of tri-diagonal matrices with prescribed eigenvalues. This aspect of the problem is discussed elsewhere[3,4,5].

4. THE NEW SETS OF EXPLICIT FORMULAS FOR BUTTERWORTH AND CHEBYCHEFF LADDER FILTERS

Referring to (6), the considerations above lead to the following new sets of explicit formulas for Butterworth ladder filters:

For all patterns we have

$$a_{vv}(x)=2S_vN_v(x)\frac{y_{vv}(x)}{y_{v-1,v-1}(x)} \quad , \qquad v=0,1,\ldots,n \tag{11}$$

where

$$\left.\begin{aligned} &S_o=1 \ , \quad S_v=\left\{\prod_{r=1}^{v} s_{2r-1}\right\}^{-1} && \text{for} \quad v=1,2,\ldots,n \\ &\qquad\qquad s_r=2\sin(r\pi/2n) && \text{for} \quad r=1,2,\ldots,2n-1 \ . \end{aligned}\right\} \tag{12}$$

The real polynomials $N_v(x)$ are divisors of $1+x^n$ or $1-x^n$ according as v is even or odd. Their form is uniquely determined by the characteristic f(u) of the pattern considered. Also,

$$N_v(0)=1 \ . \tag{13}$$

The real polynomials $y_{vv}(x)$ are generally prime to $1-x^{2n}$ and satisfy, among others, the simple relations

$$y_{vv}(0)=1 \tag{14}$$

$$y_{vv}(x)=y_{n-v-1,n-v-1}(-x) \tag{15}$$

where $y_{-1,-1}(x)$ and $y_{oo}(x)$ are defined as unity.

Turning now to particalar patterns we have

I. $N_v(x)$

for Takahasi's pattern

$$N_o(x)=1 \tag{16a}$$

$$N_v(x)=b_{v-1}(x)b_{v-3}(x)\cdot\begin{cases}\ldots b_2(x)b_o(x) & \text{for v odd} \\ \ldots b_3(x)b_1(x) & \text{for v even}\end{cases} \qquad v=1,2,\ldots n \tag{16b}$$

where

$$b_o(x)=1-x \ , \quad b_v(x)=(\varepsilon^{2v}-x)(\varepsilon^{-2v}-x) \qquad \text{for } v=1,2,\ldots,n-1 \tag{17}$$

for pattern (1)

$$N_o(x)=N_1(x)=1 \tag{18a}$$

$$N_v(x)=b_{v-1}(x)b_{v-3}(x).\begin{cases}\ldots b_3(x)b_1(x) & \text{for v even}\\ \ldots b_2(x)b_o(x) & \text{for v odd}\end{cases} \qquad v=2,3,\ldots,n-2,n \tag{18b}$$

$$N_{n-1}(x)=b_n(x)b_{n-2}(x).\begin{cases}\ldots b_3(x)b_1(x) & \text{for v even}\\ \ldots b_2(x)b_o(x) & \text{for v odd}\end{cases} \tag{18c}$$

where

$b_v(x)$ for $v=0,1,\ldots,n-1$ are the same as for Takahasi's pattern whereas $b_n(x)=1+x$.

for pattern (2)

$$N_o(x)=N_1(x)=N_2(x)=1 \tag{19a}$$

$$N_v(x)=b_{v-1}(x)b_{v-3}(x).\begin{cases}\ldots b_2(x)b_o(x) & \text{for v odd}\\ \ldots b_3(x)b_1(x) & \text{for v even}\end{cases} \qquad v=3,4,\ldots,n-3,n\,. \tag{19b}$$

$N_{n-2}(x)$ and $N_{n-1}(x)$ are similar to $N_v(x)$ except that for each of them there is a "jump in" of an additional factor; $b_{n-1}(x)$ for the first and $b_n(x)$ for the second. (Compare with (18c)).

II. $y_{vv}(x)$

for Takahasi's pattern

$$y_{vv}(x)=1 \qquad \text{for } v=-1,0,\ldots,n \tag{20}$$

for pattern (1)

$$y_{-1,-1}(x)=y_{oo}(x)=y_{n-1,n-1}(x)=y_{nn}(x)=1 \tag{21a}$$

$$y_{vv}(x)=1-\frac{c_{2v+1}}{c_1}x \qquad \text{for } v=1,2,\ldots,n-2 \tag{21b}$$

(interpolated from the complex value $y_{vv}(\epsilon^{2v+2})=-s_{2v+2}j\epsilon^{2v+1}/c_1$)

where $c_r=2\cos(r\pi/2n)$ for $r=1,2,\ldots,2n-1$ whereas s_r is the same as in (12).

for pattern (2)

$$y_{-1,-1}(x)=y_{oo}(x)=y_{n-1,n-1}(x)=y_{nn}(x)=1 \qquad (22a)$$

$$y_{11}(x)=1-\frac{c_5}{c_1}x \quad , \qquad y_{n-2,n-2}(x)=1-\frac{c_{2n-5}}{c_1}x \qquad (22b)$$

(interpolated from $y_{11}(0)=y_{n-2,n-2}(0)=1$ and $y_{11}(1)=y_{n-2,n-2}(-1)=s_1s_3$

$$y_{vv}(x)=1-\frac{s_3c_{2v+1}}{s_2}x+\frac{s_1s_3c_{2v+1}^2}{s_2^2}x^2-\frac{c_{2v-1}c_{2v+1}c_{2v+3}}{c_1^2c_3}x^3 \quad \text{for } v=2,3,\ldots,n-3 \qquad (22c)$$

(interpolated from the complex values $y_{vv}(\varepsilon^{2v+2})=s_{2v}s_{2v+2}s_{2v+4}\cdot$ $\cdot j\varepsilon^{6v+3}/c_1^2c_3$ and $y_{vv}(\varepsilon^{2v})=s_{2v-2}s_{2v}s_{2v+2}j\varepsilon^{6v+3}/c_1^2c_3$)

where c_r and s_r are the same as before.

Turning then to the Chebycheff case we have, following the notation in [1], to replace $G(s)$ and $H(s,x)$ in (1) by the polynomials $G(s,\xi)$ and $H(s,\eta)$ in s,ξ and s,η respectively, $0<\xi=k-k^{-1}$, $-\xi<\eta=h-h^{-1}<\xi$, where ξ and $|\eta|$ are the semi-minor axes of the confocal ellipses in the s-plane (with focii at ±2j) on which the zeros of $G(s,\xi)$ and $H(s,\eta)$ are located. The number of essentially different patterns for $\eta\neq 0$ is the same as that for $x\neq 0$ in the Butterworth case and will be denoted by the same notation, i.e. (i), (i,j) etc.. The form assumed by $H(s,\eta)$ for any particular pattern can readily be deduced from (34) in [1] and (4) above. Reasoning as in the Butterworth case, it is not difficult to see that for all Chebycheff patterns we have instead of (11)

$$a_{vv}(\xi,\eta)=2S_vN_v(\xi,\eta)\frac{y_{vv}(\xi,\eta)}{y_{v-1,v-1}(\xi,\eta)}, \quad v=0,1,\ldots,n \qquad (23)$$

where the real polynomials $N_v(\xi,\eta)$ in ξ,η are divisors of

$$[k^n+(-)^nk^{-n}]+[h^n+(-)^nh^{-n}] \quad \text{or} \quad [k^n+(-)^nk^{-n}]-[h^n+(-)^nh^{-n}]$$

(expressed as polynomials in ξ,η) according as v is even or odd whereas the real polynomials $y_{vv}(\xi,\eta)$ are generally prime to each of these two polynomials. Particularly, for the singlet pattern (1), the only modifications called for in the Butterworth formulas for this pattern (see (18), (17) and (21)) are:

$$b_o(x) \longrightarrow b_o(\xi,\eta) = \xi - \eta \quad , \qquad b_n(x) \longrightarrow b_n(\xi,\eta) = \xi + \eta$$

$$b_v(x) \longrightarrow b_v(\xi,\eta) = (k\varepsilon^{2v} - k^{-1}\varepsilon^{-2v} - \eta)(k\varepsilon^{-2v} - k^{-1}\varepsilon^{2v} - \eta) \quad \text{for } v=1,2,\ldots,n-1 \tag{24}$$

and

$$y_{vv}(x) \longrightarrow y_{vv}(\xi,\eta) = \frac{c_1\xi - c_{2v+1}\eta}{c_1} \qquad \text{for } v=1,2,\ldots,n-1 \tag{25}$$

REFERENCES

1. L.Weinberg and P.Slepian, "Takahasi's ressults on Chebycheff and Butterworth ladder networks", IRE Trans. on Circuit Theory, Vol. CT-8, pp. 88-101; June 1960.
2. F.R.Gantmacher, "The Theory of Matrices", Chelsea Publ. Co., New York, N.Y., Vol. 2, pp. 190-196, 1959.
3. I.Navot, "The synthesis of certain sub-classes of tri-diagonal matrices with prescribed eigenvalues", presented to the 70th Summer Meeting of the American Mathematical Society, August 1965, (abstracted in Notices, Amer. Math. Soc., Vol. 12, p.621) to appear in SIAM Journal on Applied Mathematics.
4. T.J.Marshall, Jr, "Ladder network synthesis by similarity transformations", Proc. Third Ann. Allerton Con. on Circuit and System Theory, Univ. of Illinois, Urbana, 1965, p. 200 (abstract).
5. T.J.Marshall, Jr, "Primitive matrices for doubly terminated ladder networks", to be presented at this Conference.

DESIGN OF NUMERICAL FILTERS: SCANNING FUNCTIONS AND EQUAL RIPPLE APPROXIMATION

G. E. HEYLIGER
Martin Company
Denver, Colorado

SUMMARY

Numerical or digital filters are increasingly important in the reduction, processing, and transmission of information. This report develops some additional tools for the design and use of one class of such filters. After introducing the general numerical filter, attention is focused on a particular class; namely, symmetrically weighted moving averages. It is shown that "scanning functions" are useful in designing such filters.

Several scanning functions are discussed, with considerable emphasis placed on the Chebyshev. The latter are particularly useful because of an inherent resolution-ripple trade off characteristic. The use of scanning functions is demonstrated in the design of a simple filter.

The chief features of this paper are: 1) The development and use of particular scanning functions; and 2) Derivation of alternate and approximate expressions for Chebyshev coefficients for the design of numerical filters (and incidently antenna arrays).

INTRODUCTION

Numerical filters for signal processing can be made to function as bandpass, low-pass, or shaping filters having much the same effect as conventional analog filters. The sampling theorem guarantees that for band-limited signals there will be no information loss due to the sampling.

One class of filter, the so-called moving average numerical filter, has a linear phase shift characteristic together with an arbitrary gain profile. For this class of numerical filter, the chief problem is to realize a desired gain-frequency profile through selection of the coefficients of a finite trigonometric series. In many cases, the idealized profile has one or more discontinuities; e.g., the ideal low-pass filter. The ordinary Fourier coefficients, although providing a minimum mean squared error approximation, result in a large ripple error. Other techniques, which essentially smooth the ripple in exchange for broadening the discontinuity transition region, depend on special shaping of the ideal gain function near the discontinuity. (1, 2) The question of best shape and optimum transition duration immediately arises. This paper approaches the design problem from a different aspect. The notion of a scanning function representation of a Fourier series(3) approximation is extended to a generalized scanning function. It is shown that many features of the approximation detail are determined by the character of the scanning function itself. The problem then is to study the character of scanning functions and to develop those appropriate for particular applications. This study is particularly rewarding, since the desired coefficients of the trigonometric approximation are simply related to the ordinary Fourier coefficients.

The commonly used scanning functions have shapes fixed entirely by the order of the trigonometric approximation. The class of Chebyshev scanning functions has an additional degree of freedom permitting a tradeoff between resolution and ripple amplitude.

MOVING AVERAGE FILTERS

The general linear numerical filter is obtained by taking the weighted sum of n+1 successive input samples together with the m previous output samples. If $E_o(nT)$ is the output sequence and $E_1(nT)$ the input sequence, then the output sample at time kT is given by

$$E_o(kT) = \beta_o E_1(kT) + \beta_1 E_1\left[(k-1)T\right] + \beta_2 E_1\left[(k-2)T\right] + \ldots + \beta_n E_1\left[(k-n)T\right]$$
$$- \gamma_1 E_1\left[(k-1)T\right] - \gamma_2 E_o\left[(k-2)T\right] \ldots - \gamma_m E_o\left[(k-m)T\right] \qquad (1)$$

where n, k, m are integers and T is the sampling interval. Taking the Laplace transform of this expression is straightforward, and the resulting general filter representation has been widely discussed in the literature [7] . A transfer function describing the input-output relation in the frequency domain is obtained from this operation.

$$H(s) \triangleq \frac{E_o(s)}{E_1(s)} = \frac{\beta_o + \beta_1 \epsilon^{-sT} + \beta_2 \epsilon^{-2sT} + \ldots + \beta_n \epsilon^{-nsT}}{1 + \gamma_1 \epsilon^{-sT} + \gamma_2 \epsilon^{-2sT} + \ldots + \gamma_m \epsilon^{-msT}}$$

where $s = \sigma + jw$ is the generalized frequency variable.

In the form of Eq. 1, the numerical filter is called "recursive". If none of the output samples are incorporated in Eq. 1., it is known as a "moving average" numerical filter, and the denominator of the right hand side of the transfer function H(s) is unity.

With the symmetry constraint $\beta_{n-k} = \beta_k$, and $N \triangleq \frac{n-1}{2}$ the transfer function reduces to a finite trigonometric series of the form

$$H(jw) = \epsilon^{-jNwT} \left(A_o + A_1 \cos wT + A_2 \cos 2\, wT + \ldots + A_N \cos(NwT)\right)$$
$$\text{for } s = jw \qquad (2)$$

The design problem of these filters may be simply stated: How are the A_k's best chosen in order to realize a desired gain function? Since theoretically, any arbitrary gain function profile may be realized (approximately) through a Fourier series expansion of the desired function, amplitude shaping may be accomplished with no phase distortion. For data reduction or information-processing, this feature may be invaluable. The accompanying fixed signal delay may be readily compensated or ignored. In control applications, delay may lead to instabilities, and such filters must be used with care.

SELECTION OF COEFFICIENTS

Assume that an arbitrary even gain function, G(r) is to be realized over the normalized frequency range $-1/2 \leq r \leq 1/2$, where $r \triangleq wT$. A straightforward expansion of G(r) in a Fourier series will yield a set of A_i's for Eq. 2. If G(r) is smooth and continuous, the series will be uniformly convergent and the approximation will be fairly good, depending on the number of terms taken. However, if G(r) has a discontinuity, there will be some overshoot near the discontinuity due to the nonuniform convergence of the Fourier series. This is the familiar Gibb's phenomena (see Guillemin, p 483-485).

In an effort to control or eliminate this overshoot, some workers have proposed a smooth but arbitrary transition region. (see Martin (1), Ormsby (2)). Parabolic, cosine, linear, and other cutoff functions within the transition region have been proposed. Since the width and shape of the transition region are arbitrary with this approach, the question of best shape and best transition interval arises. The resulting cutoff characteristic and overshoot depend critically on G(r) in this transition region.

The scanning function design approach is developed in the following paragraphs. This technique focuses upon the fundamental limitations of the class of filter, thus providing insight as well as a design procedure.

THE GENERAL SCANNING FUNCTION

Let G(r), the idealized gain function, be represented by the general trigonometric series

$$G(r) = \sum_{k=-\infty}^{\infty} A_k \epsilon^{j2\pi kr} \quad , \quad (-1/2 \leq r \leq 1/2)$$

For any practical application, a finite series must be used. The resulting series

$$G_N(r) = \sum_{k=-N}^{N} A_k \epsilon^{j2\pi kr}$$

is only an approximation to G(r). The difference is called the truncation error.

A particular set of coefficients, a_k,

$$a_k = \int_{1/2}^{1/2} G(u) \epsilon^{-j2\pi ku} du$$

are the familiar Fourier coefficients.

For the general set of coefficients A_k, let $A_k = b_k a_k$. Then

$$G_N(r) = \sum_{k=-N}^{N} b_k a_k \epsilon^{j2\pi kr} = \sum_{k=-N}^{N} b_k \int_{-1/2}^{1/2} G_u \epsilon^{j2\pi k(r-u)} du$$

and

$$G_N(r) = \int_{-1/2}^{1/2} G(u) \sum_{k=-N}^{N} b_k \epsilon^{j2\pi k(r-u)} du.$$

The scanning function, S(r) is defined with reference to this last expression for $G_N(r)$.

$$S(r) = \sum_{k=-N}^{N} b_k \epsilon^{j2\pi kr} \qquad (3)$$

The expression for $G_N(r)$, Eq. 4, simply states that the approximation is given by convolving the desired gain function G(r), with the scanning function S(r).

$$G_N(r) = \int_{-1/2}^{1/2} G(u) S(r-u) du. \qquad (4)$$

Note that,

$$\int_{-1/2}^{1/2} S(u) du = b_0 \text{; thus, } S(r) \text{ is normalized if } b_0 = 1$$

Important Particular Scanning Functions

If $b_0 = b_{\pm 1} = . . . b_{\pm N} = 1$, $S(r-u) \triangleq S_F(r-u)$ is the usual Fourier scanning function.

$$S_F(r) = 1 + 2\cos 2\pi r + 2\cos 4\pi r + \ldots + 2\cos 2N\pi r = \frac{\sin(2N+1)\pi r}{\sin \pi r} \qquad (5)$$

Figure 1 shows a sketch of the Fourier scanning function.

If $b_0=1$, and $b_k = \frac{N-k+1}{N+1}$, the series may also be summed as

$$S_f(r) = \frac{1}{N+1}\left\{\frac{\sin(N+1)\pi r}{\sin r}\right\}^2 \qquad (6)$$

where $S_f(r)$ is the Fejer scanning function of Figure 1. Guillemin shows that $S_f(r)$ is a Cesáro sum of the Fourier series which overcomes the non-uniform convergence problem near the discontinuity. A characteristic of this particular scanning function is that it is non-negative.

An interesting and important aspect of the scanning function concept is that once the character of this weighting function is established, the appropriate cosine filter coefficients A_k for an arbitrary gain profile may be found from the scanning function b_k and the Fourier coefficients for the gain profile $G(r)$:

$$A_k = b_k \circ a_k. \qquad (7)$$

$G_N(r)$ Characteristics in Terms of the Scanning Function

The effect of the scanning function shape is readily determined. Consider the ideal low pass filter. From graphical considerations and Eq. 4, $G_N(r)$ until the scanning function main lobe approaches the filter cutoff frequency. When the scanning function lobe is beyond cutoff, $G_N(r)$ will be near zero. Thus it is seen that the cutoff duration is approximately equal to the scanning function main lobe width. Since the lobe width in turn is directly related to the order of the polynomial for a given scanning function, the effect of taking additional terms is evident.

Continuing this discussion, let us consider a specific rectangular gain function with cutoff at ± r_1:

$$G_N(r) = \int_{1/2}^{1/2} G(u)S(r-u)du. \qquad (-1/2 \leq r \leq 1/2)$$

If $G(r)$ is to be rectangular cutoff function, then $G_N(r)$ takes on a special form

$$G_N(r) \int_{-r_1}^{r_1} S(r-u)du.$$

From this discussion it is also apparent that $S(r)$ is the important nth order approximation to the frequency delta function, or band pass filter of zero width.

The problem of choosing the A_1's has now been shifted from selection of an appropriate transition shape near r_1 in the original $G(r)$ to the general problem of selecting an appropriate scanning function. The author has employed several scanning functions derived from $S_F(u)$ that have some overshoot, but smaller amplitude side lobes. The ideal scanning function, apparently, would have a zero width main lobe and no side lobes. This ideal, of course, is not realizable with the finite trigonometric sum. Alternatives, in addition to $S_F(r)$ and $S_f(r)$, are the topic of later sections.

The search for an optimal scanning function, with a narrow main lobe and low amplitude side lobes, suggests that we borrow from the antenna designer and use what amounts to a Dolph-Chebyshev array weighting (5). This possibility is taken up in the next section

CHEBYSHEV OR EQUAL RIPPLE SCANNING FUNCTIONS

The Chevyshev scanning functions $S_c(r)$ have been singled out because of the following important properties: 1) The side lobes have the same magnitude, and they alternate in sign; 2) For a given main lobe width, the ratio of the peak lobe amplitude to the side lobe amplitude is maximized; 3) Conversely, for a given peak to side lobe amplitude ratio, the main lobe width is minimized. These properties are obviously desirable for scanning functions. In the following paragraphs, design parameters for this class of scanning functions are developed, and a derivation of the Chebyshev scanning function is outlined. Figure 1 is a graphical comparison of several scanning functions with $N = 5$.

In addition to the order, N, two additional parameters are used to describe the Chebyshev filter. We use the effective main lobe width, r_o, and the ratio of peak to side lobe amplitude, R. Figure 1 indicates the general shape of this scanning function. The cosine coefficients are obtained by 1) using a Chevyshev polynomial of order 2N to describe the equal amplitude side lobes, 2) extending the independent variable outside the range (-1,1) to obtain the desired main lobe amplitude, and 3) finally make a change of variable to obtain the results in terms of the normalized frequency variable, r. Such a development is given in Heyliger (6) and Kraus (5) (p. 103). The Chebyshev scanning function coefficients, when properly normalized are given by

$$S_c(r) = 1 + 2 \sum_{k=1}^{N} \frac{C_k}{C_0} \cos 2\pi kr \tag{8}$$

where

$$C_k \sum_{p=k}^{N} (-1)^{N-p} \frac{2Nx_o^{2p}}{N+p} \binom{N+p}{N-k} \binom{N-k}{N-p} \tag{9}$$

and $x_o = \sec \pi r_o$. While Eq. 9 is the usual form given for the coefficients in the design of linear arrays, it is particularly poor for computational purposes. For many cases of interest, neighboring terms are large and of the same order of magnitude which is disasterous with alternating signs. A superior expression can readily be obtained by expanding the C_k as a polynomial in

$$z_o^2 = x_o^2 - 1 = \tan^2 \pi r_o \tag{10}$$

This polynomial is monotonic, and is also useful in deriving approximate relationships. In terms of z_o, there obtains, after some manipulation (6)

$$S_c(r) = 1 + 2 \sum_{p=1}^{N} \mathbf{b_p} \cos 2\pi p\, r \tag{11}$$

where

$$b_p = \frac{x_o^{2p} \sum_{s=0}^{N-p-1} \frac{z_o^{2s}}{s+1} \binom{N+p+s}{s} \binom{N-p-1}{s}}{\sum_{s=0}^{N-1} \frac{z_o^{2s}}{s+1} \binom{N+s}{s} \binom{N-1}{s}} \qquad p=1,\ 2,\ .\ .\ .,\ N-1 \tag{12}$$

and

$$b_N = \frac{x_o^{2N}}{2N\, z_o^{\,2} \sum_{s=0}^{N-1} \frac{z_o^{\,2s}}{s+1} \binom{N+s}{s}\binom{N-1}{s}} \qquad (13)$$

In terms of the fundamental Chebyshev polynomial of order 2N we have the following relationships: (See Figure 2 for relation between R and r_o)

$\lvert x \rvert < 1$	$\lvert x \rvert > 1$
$T_{2N}(x) = \cos(2N\cos^{-1}x)$	$T_{2N}(x) = \cosh(2N\cosh^{-1}x)$

For the parameter R we have

$$R = \cosh(2N \cosh^{-1}R) \text{ or } x_o = \cosh\left(\frac{1}{2N} \cosh^{-1}R\right) \qquad (14)$$

From these expressions, the following useful design relationship may be derived.

$$\tan \pi r_o = \frac{1}{2}\left(Q - Q^{-1}\right) \qquad (15)$$

where

$$Q = \left[2R\sqrt{\left(\frac{1 + 1 - 1/R^2}{2} \mp \frac{1}{4R^2}\right)}\,\right]^{1/2N}.$$

Eq. 15 is exact and gives no computational difficulties for $R > 1$ and $r_o < 0.5$.

Lobe Width and Attenuation

As a review and to fix ideas at this point, it is useful to explore various relationships among N, the order of the scanning function; R, the ratio of main lobe amplitude to side lobe amplitude; and r_0, the effective half-width of the main lobe. Figure 2 summarizes these relationships graphically. We have from Eq. 15

$$\tan \pi r_0 \cong \sinh\left(\frac{1}{2N} \log 2R\right) \text{ for } R >> 1.$$

Also

$$\sinh\left(\frac{1}{2N}\log 2R\right) \cong \frac{\log 2R}{2N} + \frac{1}{3!}\left(\frac{\log 2R}{2N}\right)^3 + \ldots.$$

Now R varies between 10 and 10,000 for cases of practical interest, thus $3 < \log 2R < 10$.

The approximation

$$\tan \pi r_0 \cong \frac{\log 2R}{2N} \text{ for } \frac{\log 2R}{2N} << 1 \qquad (16)$$

is useful in many instances. Further, for $\pi r_0 << 1$, that is, for narrow lobe widths,

$$\tan \pi r_0 \cong \pi r_0 \quad \text{and } r_0 \cong \frac{\log 2R}{2\pi N}. \qquad (17)$$

In summary we have determined appropriate expressions for the design parameters of the class of Chebyshev scanning functions. The coefficients, b_p, are found to be functions of the lobe width r_0, or equivalently, the peak-to-side lobe ratio R, and the order N. Any two of the parameters r_0, R, N determine the third.

Figure 2 indicates that the approximations of Eq. 16 and 17 are quite good for design purposes. Alternatively, we have

$$R \cong 1/2\, e^{2N \tan \pi r_0},\ \left(\frac{\log 2R}{2N} << 1\right) \text{ or } R = 1/2\, e^{2N\pi r_0},\ (r_0 << 1) \qquad (18)$$

Equations 16-18 are very useful for estimating in practical applications of Chevyshev scanning functions, since they simply relate the three essential parameters; R, r_0, and N. Depending on the point of view, these approximations say that for a fixed peak to side lobe ratio R, the main lobe width r_0 decreases at 1/N. Alternately, for a fixed number of terms, N, the ratio R increases expoentially with r_0.

Binomial Scanning Function

The polynomials in x_0, which make up the Chebyshev scanning function coefficients, are dominated by the x_0^{2N} term as $R \rightarrow \infty$ as indicated in Eq. 9. In the limit, then, as $R \rightarrow \infty$, $r_0 \rightarrow \frac{1}{2}$, the C_k reduce to new coefficients, called the binomial coefficients due to the leading factor.

$$C_{n-k} = \binom{2N}{k} x_0^{2N}$$

When normalized with respect to C_0 to yeild $b_0 = 1$, these coefficients become

$$B_{N-k} = \frac{\binom{2N}{k}}{\binom{2N}{N}}$$

The resulting scanning function, also shown in Fig. 1, has the interesting feature of monotonically decreasing with r.

Scanning Function Coefficient Limiting Values

The Chebyshev scanning function coefficients are given by Fig. 12 and 13.

We have seen that as $R \rightarrow$, $r_0 \rightarrow \frac{1}{2}$, and the b_p's becomes the binomial scanning function coefficients.

$$B_p = \binom{2N}{N-p} \Big/ \binom{2N}{N}$$

For $z_0 = \tan \pi r_0 \rightarrow 0$, $r \rightarrow 1$, and the scanning function has no useful filtering properties.

As $N \rightarrow \infty$, $r_0 \rightarrow 0$, and the expression for b_p, Eq. 12. shows that for all finite p, $b_p \rightarrow 1$, i.e., the Chebyshev coefficients become the Fourier coefficients. This is a general characteristic of all the scanning function coefficients as $N \rightarrow \infty$.

Figure 3 shows how the Chebyshev coefficients vary with the change in lobe width, r_0. Since there is a practical limit on the range of coefficient values, Fig. 3 indicates that the lobe width cannot realistically be reduced to zero.

An interesting approximation to the Chebyshev b_p obtains for small z_0

$$b_p \cong x_0^{2p} \left[1-p(p+2)\,(z_0^2/2) \right] , \quad \text{for } z_0^2 << \frac{2}{N^2} ,$$

Now, $z_0^2 = \tan^2 \pi r_0$ and $x_0^2 = 1 + \tan^2 \pi r_0$.

Using the above approximation,

$$\log b_p \cong p \log (1+ \tan^2 \pi r_0) + \log \left[1 - \frac{p(p+2)}{2} \tan^2 \pi r_0 \right]$$

or

$$b_p \cong \exp \left[-p^2 \frac{\tan^2 \pi r_0}{2} \right] = b_1^{p^2} \qquad (r_0 << \sqrt{2/\pi N}) \qquad (19)$$

It is interesting to observe that the approximation $b_p = b_1^{p^2}$, derived in Eq. 19, does hold quite well.

FILTER DESIGN EXAMPLE

To demonstrate the use of the scanning function in the design of numerical filters, a specific gain function will be discussed.

Equation 4 shows that whatever the scanning function, S(r), the approximation to an idealized gain function, G(r), is given by the convolution,

$$G_N(r) = \int_{1/2}^{1/2} G(u)\, S(u-r)\, du.$$

The scanning function S(r) determines how the idealized gain function will be weighted. The resulting value of the approximation at any point r is determined by the integral of the product G(u) S(u-r). Although the major contribution to $G_N(r)$ is given by values of G near r, the values of G remote from r affect the final result according to the "window" S(r) through which G(r) is being observed. In general, the resulting $G_N(r)$ is a complicated function of the shape of the gain function and the shape of the window. Choice of S(r) does give precise control over the viewing window. It does not, in general, give precise control over the error in the approximation $G_N(r)$. Nevertheless, in some cases of interest, simple relations between the overshoot and the shapes of the gain and scanning functions can be determined. In many instances, the choice of the window shape alone goes far in placing limits on the resultant error of approximation.

Figure 4 shows four approximations to an ideal rectangular cutoff filter. These approximations for N=5 were obtained by using the Fourier, Chebyshev (R=10), Fejer, and binomial scanning functions. The following observations point up the differences among the various scanning functions. The low pass filter characteristic is simple but widely used. Note that the finite approximation to the ideal profile takes on the character of the scanning function employed. All but the Chebyshev is completely determined by the order of the filter, N. With the 10:1 ripple to peak ratio, R, the Chevyshev approximation has a slightly widertransition region than the Fourier approximation. The Fourier approximation has a sharper transition region but also has considerably more ripple in the pass and stop bands. The binomial approximation has such a broad transition region that it would be useful only for very special applications.

It should be emphasized that the approximation, Eq. 4, depends upon a convolution, i.e., the integral of the scanning function times the ideal gain profile. In general, then, the approximation will not be precisely equal ripple even though the scanning function does have this feature.

CONCLUSIONS

The main result of this report is the development of additional tools for designing numerical filters.

The general symmetrical weighted numerical filter has been shown to have a frequency function consisting of a pure time delay together with a real gain function. The delay is always equal to one half the number of equal sample periods encompassed by the filter, and the gain function may be represented as a trigonometric cosine series with arbitrary coefficients.

It has been shown that the gain function approximations of order N may be expressed as a convolution, namely the integral of the product of the idealized gain function and a scanning function, or window, through which the ideal gain function is observed. The scanning function is useful and important because much of the detail of the trigonometric approximation is independently determined by the shape of the scanning function regardless of the desired idealized gain function.

Several commonly used scanning functions are discussed. Considerable attention is devoted to the Chebyshev scanning function, since it has three cited useful properties.

Alternative expressions for the Chebyshev coefficients are derived which have good summing characteristics. In addition, simplified relations between scanning function main lobe width and side lobe amplitude are given. Approximations for the important characteristics of scanning functions are given for design estimation purposes. Such approximations are useful in thinking about and understanding the consequences of various design alternatives.

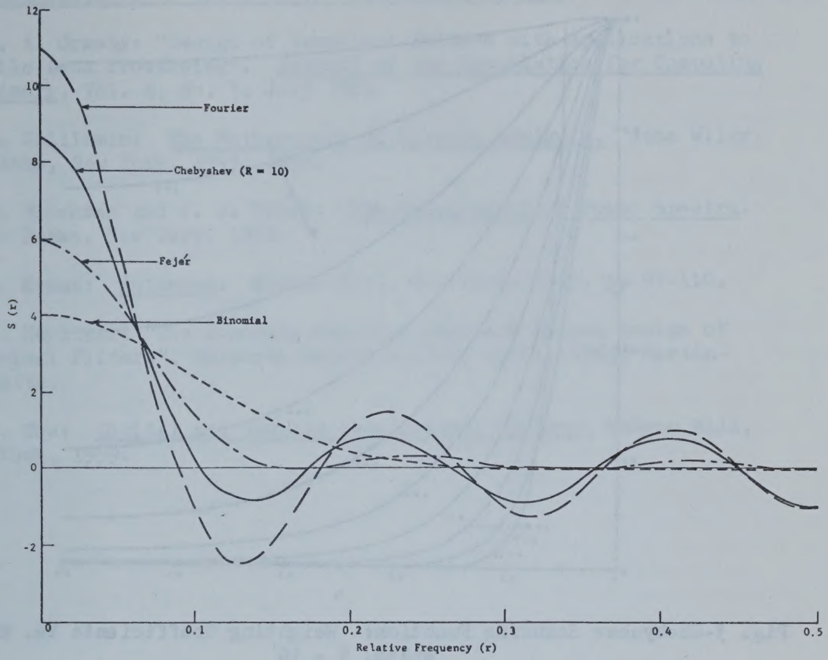


Figure 1 - Comparison of Scanning Function Characteristics,N = 5

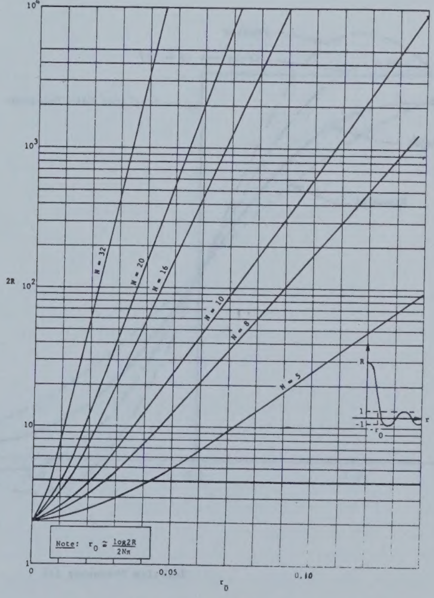


Figure 2 - Chebyshev Scanning Function: Weighting Coefficients Vs. Main Lobe Width, N=10

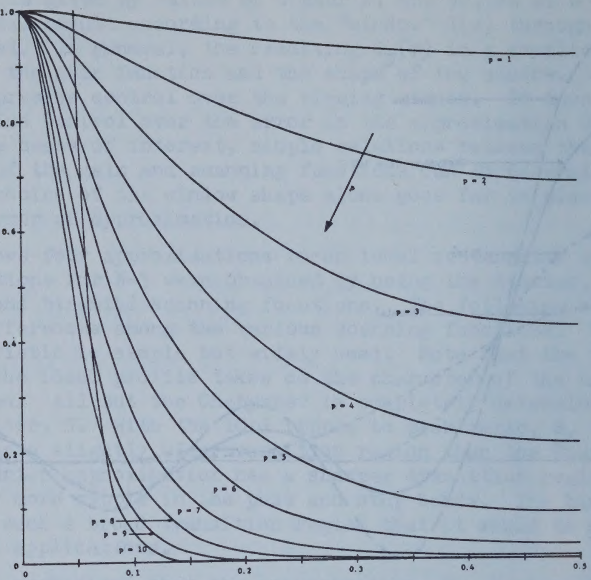


Fig. 3-Chebyshev Scanning Function: Weighting Coefficients Vs. Main Lobe Width, N = 10

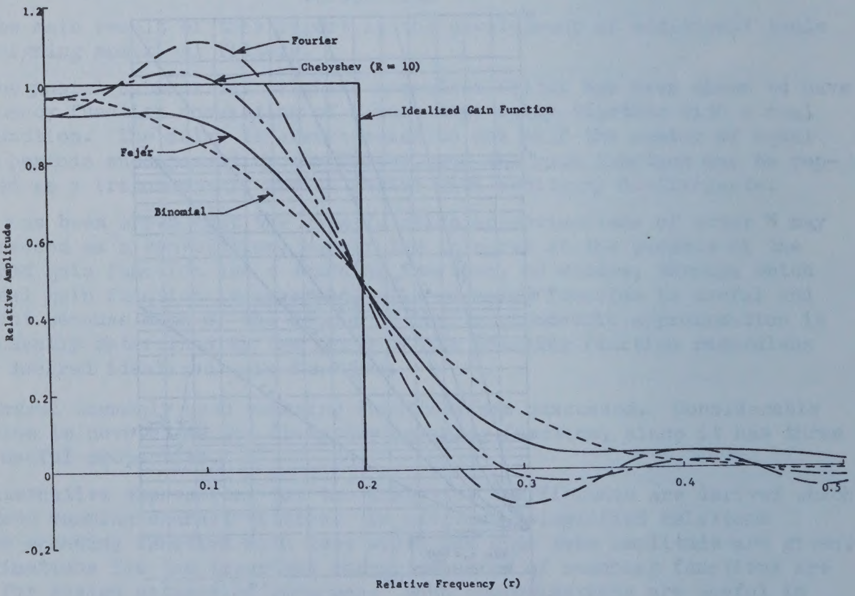


Fig. 4 - Comparative Design of Idealized Low Pass Filter, N=5

REFERENCES

1. M. A. Martin: "Frequency Domain Applications to Data Processing". IRE Transactions on Space Electronics and Telemetry, Vol. 5, March, 1959.

2. J. F. A. Ormsby: "Design of Numerical Filters with Applications to Missile Data Processing". Journal of the Association for Computing Machinery, Vol. 8, No. 3, July 1961.

3. E. A. Guillemin: The Mathematics of Circuit Analysis. John Wiley and Sons, New York, 1949, 1958.

4. R. B. Blackman and J. W. Tukey: The Measurement of Power Spectra. Dover Press, New York, 1958.

5. J. D. Kraus: Antennas. McGraw Hill, New York, 1950, pp 97-110.

6. G. E. Heyliger: "The Scanning Function Approach to the Design of Numerical Filters", Research Report R-63-2, April, 1963, Martin-Marietta.

7. J. T. Tou: Digital and Sampled Data Control Systems, McGraw Hill, New York, 1959.

EQUIRIPPLE APPROXIMATIONS OF GROUP DELAY AND ATTENUATION IN NETWORK SYNTHESIS*

R.K. KWAN
Department of Electrical Engineering
McGill University, and
Northern Electric R & D Laboratories
Ottawa, Canada.

G.G. BACH
Department of Mechanical Eng.
McGill University
Montreal, Canada.

Abstract: In this paper an approximation procedure is presented to yield a passive lossless two-port network which exhibits a pre-specified equiripple group delay and an equiripple attenuation response in the passband, and an equalminima attenuation response in the stopband(s).

The approximation method is an iterative one which begins with a realizable network function whose delay and attenuation response must possess at least a pre-determined number of variations within the frequency bands. The maxima and minima of the delay and attenuation response are then equalized through modification of the poles and zeros of the network function. The procedure used for equalization is somewhat akin to the Remez' second algorithm [1].

The delay and attenuation ripple are not independent and it was observed that the passband attenuation ripple may be lowered by reducing the pre-specified delay ripple and/or by extending the equiripple delay beyond the passband. Consequently a compromise solution may be found to satisfy a given specification.

Design examples are given to demonstrate the applicability of this approximation procedure.

INTRODUCTION

Electrical networks used for transmitting digital or video signals require a specification of not only the attenuation- but also the group delay-frequency characteristics. These requirements are usually described in the form of tolerance contours. In the passband, the variation of the group delay is to remain within a tolerance limit Dp, say, while the attenuation is to stay below a certain minimum value, Ap; in the stopband, the attenuation should be above a prescribed level A_S.

Approximation techniques have been established based on various error criteria. Methods by Fu [2] and Linvill [3] use the least square criterion. Skwirzynski and Zdunek [4] carry out the approximation in Taylor's sense. Iedokoro's technique [5] yields an equiripple attenuation and an almost flat delay characteristics. Feistel and Unbehauen [6] obtain a maximally-flat delay in the passband using Bessel polynomial and a Chebyshev stopband attenuation by properly adjusting the attenuation poles. In most approximation problems in network synthesis, however, every segment is equally important within the approximation interval and hence the "best" approximation may be regarded as one which yields a minimum equiripple error. Methods by Darlington [7], Bennett [8], and Grinich [9], have been developed to achieve this. These methods, however, require conformal mapping and are quite involved and tedious. Perhaps the most widely used procedure is the delay (phase) correction method which requires that the attenuation characteristic be approximated first and the delay subsequently equalized by means of an allpass equalizer. Unfortunately the final network obtained by this approach may require excessive circuit elements, particularly when the flat delay is to extend beyond the passband since the group delay tends to change rather abruptly near the band edge.

---

* This research was partially supported by the Northern Electric Company.

The purpose of this paper is to describe a numerical method which is capable of finding a realizable network function satisfying certain attenuation and group delay specification according to an "equiripple" criterion to be established. It has been found that this method is suitable for digital computer implementation.

THE NETWORK FUNCTION AND THE EQUIRIPPLE CRITERION

The transfer function H(p), defined as the input to output power ratio of a resistance-terminated passive lossless two-port network, may be represented by a quotient of two polynomials, i.e.

$$H(p) = \frac{E(p)}{P(p)} = \frac{h \prod_{i}^{n} (p - \lambda_{oi})}{\prod_{i}^{m} (p - \lambda_{pi})} \tag{1}$$

where p is the complex-frequency variable, h is a constant multiplier, λ_{oi} are the zeros of the function and λ_{pi} are the poles. For a realizable H(p), it is required that (i) E(p) must be a strictly Hurwitz polynomial, (ii) P(p) must be an even or odd polynomial whose degree is lower or at most equal to that of E(p), and (iii) the absolute value of H(p) for $p=j\omega$ must be equal or greater than unity.

From condition (i) we can see that there are two kinds of roots for E(p) in the left half of the p-plane, namely the real $(-\sigma_o)$ and the complex $(-\sigma_o \pm j\beta_o)$ roots. According to condition (ii) the roots of P(p) must possess the property of quadrantal symmetry taking the form of $\pm\sigma_p$ for real, $-\sigma_p \pm j\beta_p$, $\sigma_p \pm j\beta_p$ for complex and $\pm j\beta_p$ for purely imaginary roots. The imaginary roots are sometimes referred to as the attenuation poles. The group delay response $D(\omega)$ in seconds, defined as the derivative of the phase of the network with respect to the frequency, can be expressed as

$$D(\omega) = \sum_{K=1}^{n_r} \frac{-\sigma_{oK}}{\sigma_{oK}^2 + \omega^2} + \sum_{K=1}^{n_c} \frac{-2\sigma_{oK}(\sigma_{oK}^2 + \beta_{oK}^2 + \omega^2)}{(\sigma_{oK}^2 + \beta_{oK}^2 - \omega^2 + 4\sigma_{oK}^2\omega^2)} \tag{2}$$

Note that the group delay is determined exclusively by E(p). The attenuation response $A(\omega)$ in db of the network is given by

$$A(\omega) = 20 \log_{10} |H(p)|_{p=j\omega}, \tag{3}$$

or

$$A(\omega) = 10 \left\{ \log_{10} \frac{h^2 \prod_{K=1}^{n_r} (\sigma_{oK}^2 + \omega^2) \prod_{K=1}^{n_c} \left[(\sigma_{oK}^2 + \beta_{oK}^2 - \omega^2)^2 + 4\sigma_{oK}^2\omega^2 \right]}{\prod_{K=1}^{m_{r/2}} (\sigma_{pK}^2 + \omega^2)^2 \prod_{K=1}^{m_{c/2}} \left[(\sigma_{pK}^2 + \beta_{pK}^2 - \omega^2)^2 + 4\sigma_{pK}^2\omega^2 \right]^2 \prod_{K=1}^{m_{i/2}} (\beta_{pK}^2 - \omega^2)^2} \right\} \tag{4}$$

when expressed in terms of its poles and zeros. Here n_r and n_c are the respective number of real and complex zeros, and m_r, m_c and m_i are the number of real, complex and imaginary poles of H(p) respectively. Note that the quadrantal symmetry of P(p) is guaranteed by (4) in which the distinction between zeros and poles is merely a distinction between plus and minus signs. Therefore, these poles and zeros may be treated as a set of adjustable parameters, say α_i. In some cases certain attenuation poles are fixed and hence these fixed attenuation poles are not adjustable parameters. One may observe that the total number of adjustable parameters is equal to the number of zeros and half the number of poles in H(p). As an example consider the pole-and-zero pattern of a bandpass network function shown in Figure 1. This function has 12 complex zeros and 4 complex poles in the passband, 4 attenuation poles and 2 fixed attenuation poles in the stopband. The total number of adjustable parameters is therefore 16. The corresponding attenuation and delay response are shown in Figure 2 in which peaks that concave downward are identified as "maximum peaks" and those concave upward, "minimum peaks".

It is apparent that shifts in the position of these parameters will alter the attenuation and delay characteristics. The relationship between changes in these positions and changes in the network responses may be studied from the partial derivatives $\delta A(\omega)/\delta\alpha_i$ and $\delta D(\omega)/\delta\alpha_i$ which can be easily derived analytically.

Attention is now directed to the adjustment of these parameters such that (i) in the passband the delay of H(p) will exhibit an equiripple behaviour with a pre-specified delay ripple and contain a total of N_D $(=n_r+n_c+1)$ peaks, and (ii) the attenuation response should be equiripple with at least N_A $(=(m_r+m_c+2)/2)$ peaks in the passband and N_S $(=(m_i/2)+1)$ minima in the stopband. This requirement constitutes what we call the "equiripple criterion".

THE APPROXIMATION PROCEDURE

The approximation procedure can be stated in four basic steps: (i) The choice of a realizable network function whose delay and attenuation response must possess the required number of peaks; i.e. there are N_D delay and N_A attenuation peaks in the passband and N_S attenuation minima in the stopband. (ii) The identification of extrema in the delay and attenuation response. (iii) The equalization of these extrema by the adjustment of poles and zeros. (iv) Repeat steps (ii) and (iii) until the equiripple criterion is satisfied. The details of each of these steps are now discussed.

(i) Initial Choice of a Network Function

It is necessary that the initial pole and zero positions of a network function be properly chosen to meet the requirements of peaks according to the equiripple criterion. Certain amount of discretion should be exercised by the designer in making such a selection. In general an initial network function may be chosen on the basis of known solutions to problems which are similar to the problem under investigation, or by placing the complex and real poles and zeros sufficiently close to the jω-axis within the passband and all the attenuation poles within the stopband. A typical arrangement for the latter approach can be referred to Figure 1.

(ii) Identification of Extrema

Peaks are identified according to their magnitudes, i.e. the highest or most extreme one is chosen first. If the required number is even, half the peaks chosen will be the maximum peaks and the other half, minimum peaks. If the number is odd, the last peak selected could be either a maximum or a minimum peak depending on whether it is closer in amplitude to the last maximum or the last minimum peak already chosen.

(iii) Equalization of Peaks

The peaks in the respective delay and attenuation response are equalized by adjusting the parameters. The amount of parameter changes, $\Delta\alpha_i$, are computed according to an iterative procedure formulated in the following manner.

At the r-th iteration let the magnitude of those peaks in the respective delay and attenuation response in the passband be identified as:

$$D_k^{(r)} = D(\omega_k^{(r)}) \qquad (k=1,2,\ldots N_D) \tag{5a}$$

and

$$A_k^{(r)} = A(\omega_k^{(r)}), \qquad (k=1,2,\ldots N_A) \tag{5b}$$

and the attenuation minima in the stopband are

$$A_{sk}^{(r)} = A(\omega_k^{(r)}) \qquad (k=1,2,\ldots N_S) \tag{5c}$$

for $r = 1,2,\ldots.,$ where ω_k are the frequencies corresponding to the extrema in the various bands. The iterative scheme is to determine a set of $\Delta\alpha_i$ such that in the passband the delay peaks will be equiripple about an average delay level L_D with a pre-specified delay ripple S_D, and the attenuation peaks will also be equiripple about an attenuation level L_A with an unknown attenuation ripple S_A. In addition, all the attenuation minima in the stopband are equal to an unknown constant level A_{min}. According to first-order approximation, one may establish the following equations, i.e.

$$D_K^{(r)} + \sum_{i=1}^{N_A+N_D-3} \left(\frac{\delta D_K^{(r)}}{\delta\alpha_i}\right) \Delta\alpha_i^{(r)} = L_D + (-)^{N_D+K+1} (S_D/2) \qquad (K=1,2,\cdots\cdots,N_D) \tag{6a}$$

$$A_K^{(r)} + \sum_{i=1}^{N_A+N_D-3} \left(\frac{\delta A_K^{(r)}}{\delta\alpha_i}\right) \Delta\alpha_i^{(r)} = L_A \pm (S_A/2) \qquad (K=1,2,\cdots\cdots,N_A) \tag{6b}$$

$$A_{SK}^{(r)} + \sum_{i=1}^{N_A+N_D-3} \frac{\delta A_{SK}^{(r)}}{\delta\alpha_i} \Delta\alpha_i^{(r)} = A_{min} \qquad (K=1,2,\cdots\cdots,N_S) \tag{6c}$$

The term $(-)^{N_D+k+1}(S_D/2)$ in (6a) indicates that the last delay peak must be a minimum peak because the group delay tends to be zero toward high frequency. The "+" or "-" sign of $S_A/2$ in (6b) corresponds to the maximum or the minimum attenuation peak chosen. Eliminating L_D in (6a), L_A in (6b) and A_{min} in (6c), and arranging to have all the unknowns in the left-hand side, one has

$$\sum_{i=1}^{N_A+N_D-3} \left(\frac{\delta D_K^{(r)}}{\delta\alpha_i} - \frac{\delta D_{K+1}^{(r)}}{\delta\alpha_i}\right) \Delta\alpha_i^{(r)} = D_{K+1}^{(r)} - D_K^{(r)} + (-)^{N_D+K-1} S_D \qquad (K=1,2,\cdots\cdots,N_D-1) \tag{7a}$$

$$\sum_{i=1}^{N_A+N_D-3} \left(\frac{\delta A_K^{(r)}}{\delta\alpha_i} - \frac{\delta A_{K+1}^{(r)}}{\delta\alpha_i}\right) \Delta\alpha_i^{(r)} + (\pm S_A, 0) = A_{K+1}^{(r)} - A_K^{(r)} \qquad (K=1,2,\cdots\cdots,N_A-1) \tag{7b}$$

$$\sum_{i=1}^{N_A+N_D-3} \left(\frac{\delta A_{SK}^{(r)}}{\delta\alpha_i} - \frac{\delta A_{SK+1}^{(r)}}{\delta\alpha_i}\right) \Delta\alpha_i^{(r)} = A_{S,K+1}^{(r)} - A_{SK}^{(r)} \qquad (K=1,2,\cdots\cdots,N_S-1) \tag{7c}$$

where $(\pm S_A, 0)$ denotes $+S_A$ $-S_A$, or 0 which depends on the relative adjacent attenuation peak locations. The solution of (7) gives a set of parameter changes $\Delta\alpha_i^{(r)}$ and a corresponding attenuation ripple S_A. The new network function for the next iteration will now become

$$H^{(r+1)}(\alpha_i, p) = H^{(r)}(\alpha_i^{(r)} + \Delta\alpha_i^{(r)}, p) \tag{8}$$

The process is repeated until the equiripple criterion is satisfied. The constant multiplier h is the last to adjust so that

$$|H(p=j\omega)|_{min} = 1 . \tag{9}$$

It should be noted that once the group delay is equiripple, adjustments of zeros of H(p) are no longer needed and the equalization process is then reduced to solving (7b) and (7c). It has been found that the equalization process may also be carried out in two separate stages - the first is to solve (7a) and (7b) with all the attenuation poles fixed until equiripple delay is achieved and then solve (7b) and (7c) with E(p) fixed. It is apparent that the approximation of either group delay or attenuation is merely a special case of this procedure.

Since S_A and A_{min} are not specified beforehand, the final solution may fall into one of the following three catagories:

(i) $S_D \leq D_p$, $S_A \leq A_p$ and $A_{min} \geq A_s$.

Solution is acceptable. The degree of the network function may be reduced if there is excessive stopband attenuation.

(ii) $S_D \leq D_p$, $S_A > A_p$ and $A_{min} < A_s$.

Solution is not acceptable since the resultant S_A and A_{min} do not satisfy the specification. It is quite likely that the order of the network chosen is too low.

(iii) $S_D \leq D_p$, $S_A > A_p$ and $A_{min} > A_s$.

Solution does not meet the specification as it stands now because S_A is too large. This S_A may be reduced by either decreasing the pre-specified delay ripple S_D or extending the equiripple delay beyond the passband. However, the attenuation in the stopband will be lower as well. This trade-off situation may be explained by the fact that the magnitude of delay and attenuation ripple is directly related to the pole and zero positions of H(p) in the complex plane. The closer these singularities are to the jω-axis, the larger is the ripple. Hence the zeros of E(p) will drift away from the jω-axis if the specified delay ripple is decreased or the delay passband is widened up. As a result the associated passband attenuation response becomes smoother. Since P(p) is used to shape the attenuation response associated with E(p) for equiripple behaviour, one can see that the smoother the attenuation response associated with E(p), the lower the stopband attenuation, and the smaller is the passband attenuation ripple. The amount of delay ripple reduction and how far the equiripple delay should be extended beyond the passband may be determined by trial and error.

One should realize that the magnitude of the parameter changes, $|\Delta\alpha_i|$, as computed from (7) has to be small. The reason is two-fold: (i) Equation (7) is formulated according to first-order approximation which is valid only if $|\Delta\alpha_i|$ are small, (ii) the zeros of H(p) must remain in the left-half of the complex plane because of the realizability condition. Hence the validity of these calculated $\Delta\alpha_i$ must be verified in each iterative step. To ensure realizability one may restrict the fractional parameters changes $|\Delta\alpha_i/\alpha_i|$ to be less than a number γ, say, where $0<\gamma<1$. If the largest fractional parameter change for any one type of network singularities (complex zeros, real zeros, complex poles, etc.) is found to be greater than γ, allow this particular parameter a maximum change of γ, and re-scale the rest of the parameter changes of the same type down proportionally.

To ensure the validity of first-order approximation, one may observe that if the delay response yields the correct ripple S_D, the values of the right-hand side of (7a) should be zero. It has been found that the sum of the square of these values is a good measure of how close a particular delay response is from the optimum. Accordingly if the sum in one iteration is found to be less than that in the previous cycle, and at the same time both the delay and the attenuation response maintain the required number of peaks, then the new network function is considered to be "improved". If not, reduce γ by half until the "improved" condition is reached. This may be termed the "halving" process. When the delay is equiripple, the maximum absolute attenuation deviation may be used instead to check the improved condition.

GENERAL COMMENTS

We have described an approximation method which can be used to obtain a passive lossless two-port network function satisfying a given attenuation and group delay specification according to the equiripple criterion. Fundamentally this criterion makes full use of E(p) for the approximation of group delay, and P(p) for attenuation. Whether or not a solution can always be found depends on the existence of the delay approximation. Although there is not as yet any rigorous proof for the existence and uniqueness of the equiripple delay approximation, solutions have been obtained for all the cases tried. This concurs with the viewpoint of Ulbrich and Piloty [10] who strongly suspect that solutions for equiripple delay approximation do uniquely exist. Once a suitable E(p) is found, the attenuation approximation is then reduced to a polynomial approximation problem. The solution obtained can now be considered to be "optimal" only in the sense that it satisfies the equiripple criterion and meets the specification for a network function chosen.

Mathematically the procedure described here is somewhat similar to the Remez second algorithm which has found many useful applications in rational Chebyshev approximations. This technique differs from the Remez in two aspects: (i) it adjusts the poles and zeros instead of the coefficients of the network function and hence realizability conditions can always be guaranteed. (ii) it solves a set of non-linear equations through the use of first-order approximations.

Since the iterative procedure is established on the basis of first-order approximation, it will converge if the linear approximation holds. This, however, varies from problem to problem and is very difficult to predict ahead of time. Nevertheless, it is known that if the initial function is sufficiently close to the final solution, the necessary parameter changes are small. Hence the second-and higher-order effects are negligible and the linear approximation is valid. Since the initial network function may well be poor to begin with, the rate of convergence of this procedure could be slow or may never converge at all. One remedy to this is to try another network function. The establishment of the "improved" condition coupled with the "halving" process provides a simple and systematic way to achieve such a new network function. It should be pointed out that the halving process does not guarantee convergence. In fact, it has been encountered in practice that during equalization while the delay response was improving, the attenuation response was not able to maintain the required number of peaks and the repeated application of halving process did not improve the situation. This can be explained by the fact that the improved condition is estiblished based on the equations associated with delay only. To overcome this it is usually just a matter of modifying the real part of some of the roots of P(p). The particular root(s) to be modified can be easily identified from the attenuation response. In general we have found that if the delay and attenuation response maintain the required number of peaks for every iterative cycle, the procedure will converge.

Since the incremental parameter changes are computed from the matrix given in (7), the higher the order of the network function, the larger is the matrix. In order to maintain solution accuracy, the matrix inversion routine and the values of the partial derivatives must be accurate. The Gauss-Jordan method used for matrix inversion was found to be quite satisfactory. Since the complex and real poles are used to shape the passband attenuation, a non-minimum phase network results. These complex transmission zeros may be realized by Guillemin's cascade synthesis technique[11].

RESULTS AND DISCUSSIONS

The proposed approximation procedure has been programmed on the IBM 7044 installation at McGill University. This program has been tested with a number of filter-design problems and two of these results are presented here.

Example 1. Design a lowpass network to meet the following specifications:
passband delay ripple $D_p \leq 0.05$ sec at $\omega_o = 1$ rps, over $0 \leq \omega \leq 0.7$
passband attenuation ripple $A_p \leq .2$ db, over $0 \leq \omega \leq 0.4$
stopband attenuation level $A_s \geq 63$ db, over $1.013 \leq \omega \leq \infty$

Solution: The network function H(p) was chosen to contain 10 complex zeros, 4 complex poles and 4 attenuation poles making a total of 14 adjustable parameters. The frequency response and the realization of the final network function are given in Figure 3. The unit impulse response of this final network is compared with the ideal one as shown in Figure 4.

Example 2. Design a bandpass filter to meet the following specifications:
passband delay ripple $D_p \leq 100$ µs at $f_o = 1955$ cps, over $1916 \leq f \leq 1994$,
passband attenuation ripple $A_p \leq 1.3$ db, over $1916 \leq f \leq 1994$,
lower stopband attenuation level $A_{-s} \geq 50$ db, over $0 \leq f \leq 1820$
upper stopband attenuation level $A_s \geq 50$ db, over $2090 \leq f \leq \infty$

Solution: The network function H(p) was chosen to contain 12 complex zeros, 4 complex poles, 4 attenuation poles and 2 fixed attenuation poles at zero frequency making a total of 16 adjustable parameters. The center frequency $f_o = 1995$ cps was normalized to $\omega_o = 1$ rps. This gives a bandwidth of 0.04 rps. The corresponding delay ripple at $\omega_o = 1$ rps is now 1.23 sec. Since the bandwidth is too narrow, the frequency axis was multiplied by a factor of 10 during computation.

Two sets of runs with various pre-specified delay ripple S_D were obtained and tabulated in Table 1 in which Category A maintains the same passband for both group delay and attenuation while catagory B extends the equiripple delay to cover a wider frequency range, for $.97 \leq \omega \leq 1.03$. It can be seen that solutions from A1 to A5 are not acceptable because the passband attenuatuion ripple is too large. The result of B1 also satisfies the specification but the solution of B2 is not acceptable because the stopband attenuation level is too low. The normalized frequency response and the realization of A6 are given in Figure 5. The unit impluse response of A6 is compared with the ideal one as shown in Figure 6.

From Table 1 we see that the solutions from A6 to A8 are perfectly satisfactory as far as the specification is concerned. This indicates that the network function which satisfies the specification may not be unique. The choice of the final solution should therefore be dictated by other practical considerations such as the impulse response of the network function, the configuration and the element values of the resultant network, etc.

For comparison purposes Example 1 was re-designed by the delay equalizer approach. It was found that the use of the 5th-degree E(p) and a 2nd-degree P(p) was sufficient to meet the attenuation specification. The corresponding delay deviation over $0 \leq \omega \leq 0.7$ was 0.52 sec. With the addition of a 5-section all-pass delay equalizer, the delay deviation was reduced to 0.0795 sec. Although the overall order of E(p) and P(p) has now reached 15-th and 12-th degree respectively, this network does not as yet satisfy the delay specification which requires $D_p \leq 0.05$ sec. Based on similar results of other test examples, we believe that the proposed technique has definite advantage over the delay correction method in this respect.

CONCLUSIONS

A simple iterative procedure has been developed to carry out approximations of group delay and attenuation for a reactive two-port through the adjustments of the poles and zeros of a chosen network configuration. This procedure can be easily programmed on the digital computer. The existence and uniqueness of the solution and the convergence of this technique have been discussed. Practical design examples are given. It is hoped that this method will be useful in filter design.

ACKNOWLEDGMENT

The first-named author wishes to express his sincere thanks to Dr. W.F. McGee for his helpful discussions and suggestions; to Mr. J.D. MacDonald for his comments, to Messrs M. Reshef and N. Jodoin for their assistances in network realization, and to Mrs. U.B. Manley for writing the computer programs.

REFERENCES

1. Ralston, A. "Rational Chebyshev Approximation by Remez' Algorithm", Numerische Mathematik 7, pp. 322-330; 1965.
2. Fu, K.S., "Least Squares Approximation for both Magnitude and Phase by Rational Functions", Tech. Note No. 6, Engineering Experiment Station, University of Illinois, Urbana, Illinois; February 1959.
3. Linvill, J.G., "The Approximation with Rational Function of Prescribed Magnitude and Phase Characteristics", Proc. IRE, Vol. 40, pp. 711-721; June 1952.
4. Skwirzynski, J.K. and Zdunek, J., "Design of Networks with Prescribed Delay and Amplitude Characteristics", The Marconi Review, pp. 115-139; 3rd Quarter 1960.
5. Iedokoro T., Ishizaki, Y. and Watanabe, H., "On the Approximation and Synthesis of Tchebycheff Filters with Flat Group Delay Characteristics" Proceedings First Annual Allerton Conference on Circuit and System Theory, Vol. 1, pp. 16-28; November 1963.
6. Feistel, V.K.H. and Unbehauen, R., "Tiefpasse mit Tschebyscheff-Charakter der Betriebsdampfung im Sperrbereich and maximal gebneter Laufzeit", Frequency, No. 8, Vol. 19, pp. 265-282; 1965.
7. Darlington, S. "Network Synthesis Using Tchebycheff Polynomial Series", BSTJ, Vol. 31, pp. 613-665, July 1952.
8. Bennett, B.J., "Synthesis of Electric Filters with Arbitrary Phase Characteristics", IRE Convention Record, Part 5, pp. 19-26, 1953.
9. Grinich, V.H., "Approximating Band Pass Attenuation and Phase Functions" IRE Convention Record, Part 2, pp. 96-104; 1954.
10. Ulbrich, E. and Piloty, H., "Über Den Entwurf Von Allpassen, Tiefpassen und Bandpassen mit einer im Tschebycheffschen Sinne Approximierten Kinstanten Gruppenlaufzeit", Archiv der Elektrischen Übertragung,

Band 14, No. 10, pp. 451-467; October 1960.
11. Guillemin, E.A., "Synthesis of Passive Networks", John Wiley and Sons Inc., 1957.

TABLE 1. VARIOUS RUNS FOR EXAMPLE NO. 2

RUNS	DELAY RIPPLE S_D		ATT RIPPLE S_A(db)	STOPBAND ATT LEVEL A_{min}(db)	
	$.98 \leq \omega \leq 1.02$ $\omega_o = 1$rps	$1916 \leq f \leq 1994$ $f_o = 1955$cps	$.98 \leq \omega \leq 1.02$ OR $1916 \leq f \leq 1994$	$0 \leq \omega \leq .931$ OR $0 \leq f \leq 1820$	$1.069 \leq \omega \leq \alpha$ $2090 \leq f \leq \alpha$
A1	1.23 SEC	100 µS	2.32	64.32	62.31
A2	.35 SEC	28.4 µS	1.63	60.56	58.39
A3	.30 SEC	25 µS	1.56	59.93	57.92
A4	.25 SEC	20 µS	1.48	59.36	57.36
A5	.20 SEC	16.3 µS	1.39	58.85	56.69
A6	.15 SEC	12.2 µS	1.30	57.36	55.60
A7	.10 SEC	8.1 µS	1.15	56.75	54.60
A8	.05 SEC	4.05 µS	0.95	54.52	52.53
* B1	1.23 SEC	100 µS	0.5	52.66	50.67
* B2	.15 SEC	12.2 µS	0.314	44.84	43.32

* EQUIRIPPLE GROUP DELAY IS EXTENDED OVER THE FREQUENCY RANGE FOR $.97 \leq \omega \leq 1.03$ OR $1896 \leq f \leq 2014$.

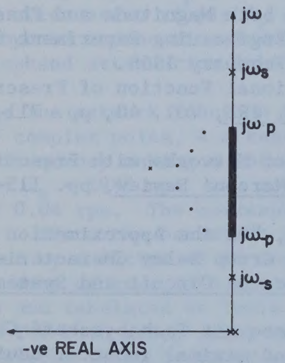


FIG. I INITIAL POLE-AND-ZERO PATTERN OF A BANDPASS NETWORK.

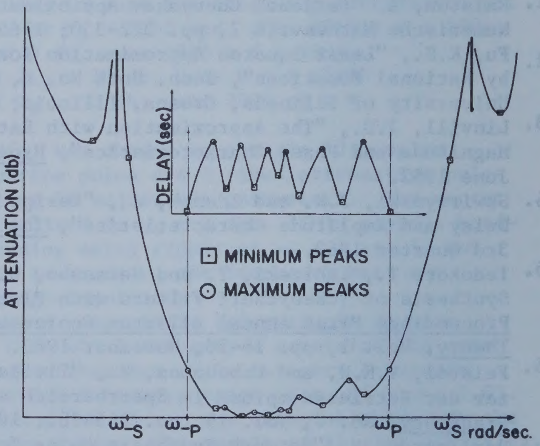


FIG. 2 FREQUENCY RESPONSE OF FIG. I

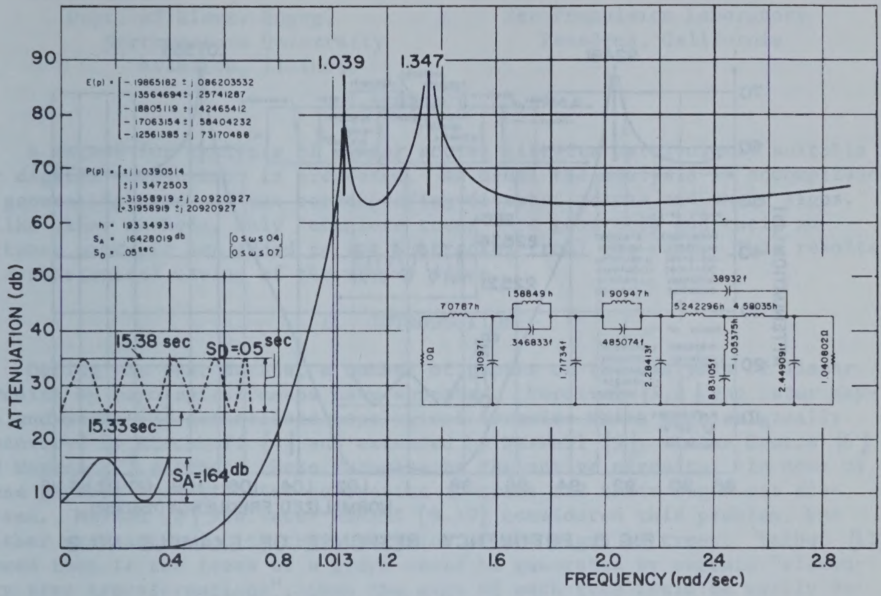


FIG. 3 FREQUENCY RESPONSE OF EXAMPLE NO. I

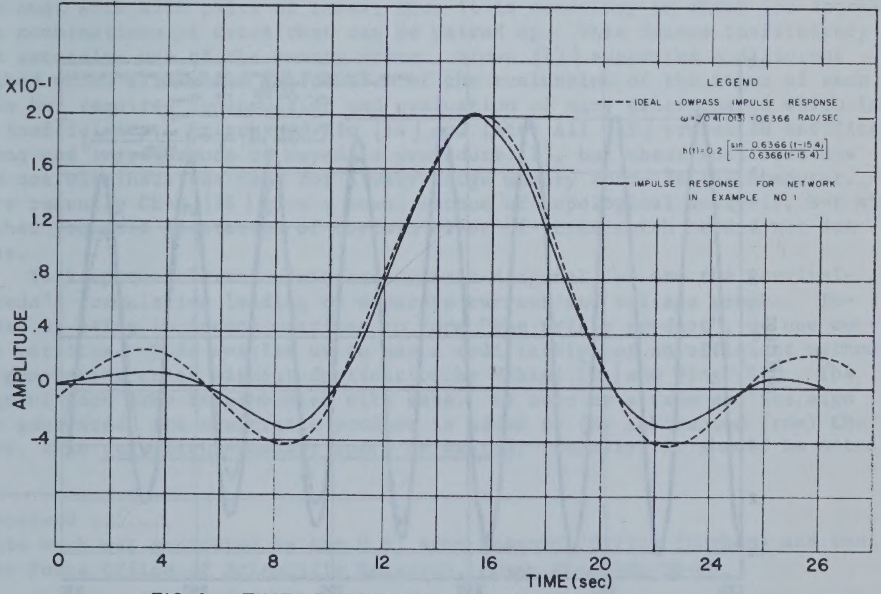


FIG. 4 TIME RESPONSE OF EXAMPLE NO. I

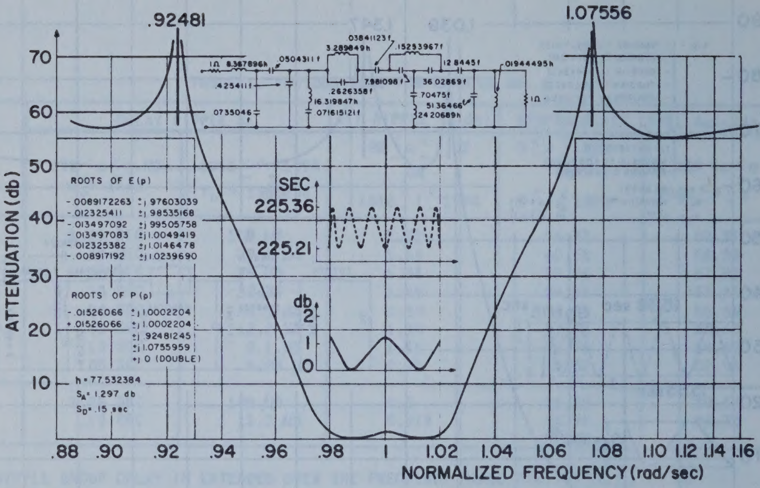


FIG. 5 FREQUENCY RESPONSE OF EXAMPLE NO. 2

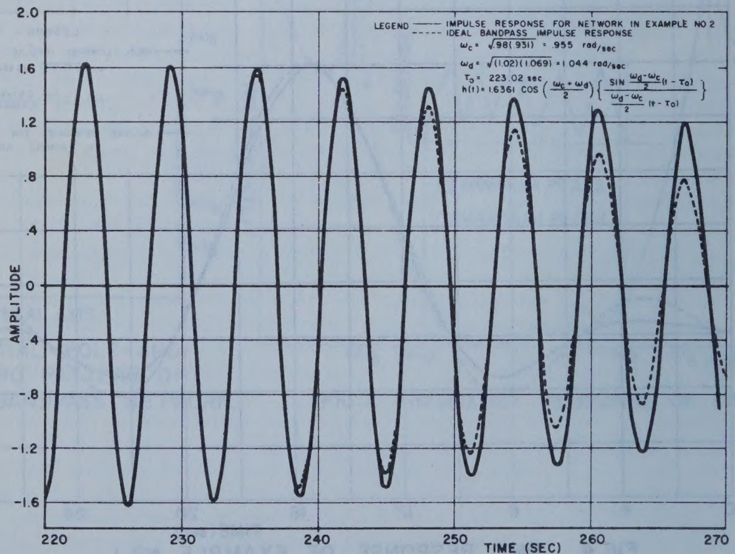


FIG.6 TIME RESPONSE OF EXAMPLE NO. 2

A TOPOLOGICAL APPROACH TO THE ANALYSIS OF LINEAR CIRCUITS[*]

S. L. Hakimi
Dept. of Elect. Engrg.
Northwestern University
Evanston, Illinois

N. Deo[+]
Jet Propulsion Laboratory
Pasadena, California

ABSTRACT

A method for analysis of linear active circuits particularly suitable for digital computation is presented. As usual the analysis is accomplished by generating trees of the corresponding directed graphs and their signs. Unlike other methods, only "complete trees" are generated and their admittance products are added to (or subtracted from) the rest. This results in a substantial saving of the memory space.

I. INTRODUCTION

During the past decade, a number of papers on the analysis of linear circuits by topological means have appeared. Percival [1,2] and later Mayeda and Seshu [3] generalized topological formulas which were originally discovered by Kirchhoff [4] and extended by Maxwell [5]. Later Coates [6] and Mayeda [7] extended these formulas to the active circuits. In none of these papers, the efficient generation of trees and their signs was discussed. Mayeda [8] and later Hakimi [9,10] considered this problem, but neither paid attention to the generation of the signs of trees. Talbot [11] showed that if the trees of a graph could be generated by certain "elementary tree transformations", then the sign of each tree could be easily determined. Mayeda and Seshu [12] showed that trees could be generated systematically by "elementary transformations", hence seemingly made Talbot's idea practical. However, a closer look showed that in using Talbot's idea, one must work with pairs of trees; thus it is necessary to check (or store) the combinations of trees that can be paired up. This causes inefficiency and excessive use of the memory space. Brown [13] suggested a different method which eliminated the problem of the evaluation of the signs of each tree but required construction and evaluation of many determinants which led to inefficiency. Frisch and Kim [14] and later Ali [15] presented modifications and improvements of Mayeda's procedure [7], but these modifications did not eliminate the need for a very large memory space in the computer. More recently Chen [16] gave a novel method of topological analysis, but his method requires generation of certain class of trees which he did not discuss.

To keep the element admittance matrix diagonal, we use the Percival-Mayeda's formulation leading to separate current and voltage graphs. Instead of using incidence matrices to form "the triple product", we use cut-set matrices. This enables us to use a modification of an efficient method of generating trees without duplication by Hakimi [9] and Wing [17]. The sign of each tree is generated with ease. As soon as a tree and its sign are generated, its admittance product is added to (or subtracted from) the rest, thus very little memory space is wasted. Finally, it should be noted

---

[*] Received ......
This work was supported by the U.S. Army Research Office (Durham) and the Air Force Office of Scientific Research, grant AF-AFOSR-98-65.

[+] Formerly with Department of Electrical Engineering, Northwestern University, Evanston, Illinois.

that one could certainly analyze a circuit by non-topological means; but, we feel, the topological method is the best when one is interested in "literal evaluation" of the network functions.

II. A MODIFICATION OF PERCIVAL-MAYEDA FORMULATION

Throughout this paper, we will concentrate on finding (the Laplace Transform of) the voltage between a pair of nodes due to a single current source, assuming that all of the initial conditions are zero. The reader can convince himself that no generality is lost by these assumptions. For example, if the source is a voltage source, it can be replaced by its Norton equivalent current source.[2]

Following Percival, we describe the voltage-current relationship for every circuit element (branch) as follows

$$y_{pq,mn} V_{pq} = I_{mn} \tag{1}$$

where V_{pq} is the voltage across a branch connected between vertices (nodes) p and q (with the positive polarity at p), and I_{mn} is the current through a branch connected between nodes m and n (with positive directions of flow from m to n) due to V_{pq}, and $y_{mn,pq}$ is the transadmittance. For an R, L, or C branch, p = m and q = n and the transadmittance becomes the admittance. For a vacuum tube V_{pq} is the voltage between the grid and the cathode and I_{mn} is the value of the controlled current source. For further discussion of these assumptions and their implications the reader is referred to Seshu and Reed [18].

From a linear active n-node circuit, we construct two directed n-vertex graphs, a voltage graph G_v and a current graph G_i. We pick two sets of n points to be associated with vertices of G_v and G_i and number these vertices 1 through n. The vertex v_i in both graphs correspond to node n_i in the circuit. For every circuit element y_j whose voltage-current relationship is governed by Eq. (1), we draw a branch, labeled b_j, in G_v and G_i. Branch b_j in G_v is connected between vertices v_q and v_p with arrowhead toward v_p, and b_j in G_i is connected between vertices v_m and v_n with arrowhead toward v_n. The current source is a branch in the current graph, whose corresponding branch in the voltage graph is assumed to be the desired voltage (see the example).

The two graphs have the same rank, R, and the same number of branches, N. Without the loss of generality, let us assume the set of branches b_1, $b_2, \ldots, b_R$ form a tree, t, in G_v and in G_i. We assume, for the moment, that branch b_1 corresponds to the current source in G_i and the voltage response in G_v. Note that branches could be renumbered if necessary. Let Q_i and Q_v be the RxN fundamental cut-set matrices of G_i and G_v with respect to tree t, and the k-th column ($1 \leq k \leq N$) in Q_i and Q_v is assumed to correspond to branch b_k [18]. I_b and V_b are column vectors of N components, where the k-th components of I_b (or V_b) correspond to the current through (or voltage across) branch b_k in G_i (or G_v). Y is a diagonal NxN matrix such that $YV_b = I_b$. Note that the diagonal entries of Y (except y_1) are all known constants or known functions of complex variables. The Kirchhoff's current law requires that $Q_i I_b = 0$. If the first R components of V_b is denoted by vector V_t, then it is known that $V_b = Q_v^T V_t$ (where Q_v^T is the transpose of Q_v). Combining the above three equations in the usual manner, we obtain

---

[2] If the voltage source has no impedance in series with it, one can assume a resistance R_g to be a series with it, then replace the combination by its Norton equivalent, once the desired function is computed. the limit as $R_g \to 0$ is evaluated. For details see the example in Sec. IV.

$$Q_i \; Y \; Q_v^T \; V_t = 0 \tag{2}$$

For Eq.(2) to have a non-zero solution, we must have $|Q_i \; Y \; Q_v^T| = 0$. Since $|Q_i \; Y \; Q_v^T| = y_1 \Delta_1 + \Delta_o$, we have $y_1 = -\Delta_o / \Delta_1$. Therefore, the problem is reduced to efficient computation of $|Q_i \; Y \; Q_v^T|$.

III. EVALUATION OF $|Q_i \; Y \; Q_v^T|$

Let I be the set of all R integers out of N; more precisely let $I = \{(i_1 i_2 \ldots i_R) \mid 1 \le i_1 < i_2 < i_3 \ldots < i_R \le N\}$. Let $Q_i(i_1 i_2 \ldots i_R)$ be an RxR submatrix of Q_i consisting of columns $i_1 i_2 \ldots i_R$ of Q_i. Define $Q_v(i_1 \ldots i_R)$ similarly. Since Y is diagonal, according to Cauchy-Binet's Theorem.

$$|Q_i \; Y \; Q_v^T| = \sum_{(i_1 \ldots i_R) \in I} y_{i_1} y_{i_2} \ldots y_{i_R} \; |Q_i(i_1 i_2 \ldots i_R)| \cdot |Q_v(i_1 i_2 \ldots i_R) \tag{3}$$

where $y_{i_1} y_{i_2} \ldots y_{i_R}$ is the product of the admittances (or transadmittances) of branches $b_{i_1} \; b_{i_2} \ldots b_{i_R}$. It is known that $|Q_i(i_1 i_2 \ldots i_R)| \, |Q_v(i_1 i_2 \ldots i_R)| = \pm 1$ if and only if branches $b_{i_1} \; b_{i_2} \ldots b_{i_R}$ form a tree of G_i and G_v [18]; and otherwise $|Q_i(i_1 \ldots i_R)| \, |Q_v(i_1 \ldots i_R)| = 0$. Therefore, the terms that contribute to the sum in Eq. (3) are those admittance products that correspond to trees in both G_i and G_v. The sum in Eq. (3) can be written, as in Eq.(4), such that the first term is due to t, second term is due to all trees of distance one from t, the third term is due to all trees of distance two from t, and so forth. The distance between tree t_j and t is the number of branches in t_j that is not in t [9].

$$\begin{aligned}
|Q_i Y \; Q_v^T| = y_1 y_2 \ldots y_R &+ \sum_{\substack{(i_1 \ldots i_R) \in I \\ i_{R-1} \le R < i_R}} y_{i_1} \; y_{i_2} \ldots y_{i_R} \; |Q_i(i_1 \ldots i_R)| \cdot |Q_v(i_1 \ldots i_R) \\
&+ \sum_{\substack{(i_1 \ldots i_R) \in I \\ i_{R-2} \le R < i_{R-1}}} y_{i_1} \; y_{i_2} \ldots y_{i_R} \; |Q_i(i_1 \ldots i_R)| \cdot |Q_v(i_1 \ldots i_R)| \\
&+ \ldots + \sum_{\substack{(i_1 \ldots i_R) \in I \\ i_1 \le R < i_2}} y_{i_1} \; y_{i_2} \ldots y_{i_R} \; |Q_i(i_1 \ldots i_R)| \cdot |Q_v(i_1 \ldots i_R)| \\
&+ \sum_{\substack{(i_1 \ldots i_R) \in I \\ R < i_1}} y_{i_1} \; y_{i_2} \ldots y_{i_R} \; |Q_i(i_1 \ldots i_R)| \cdot |Q_v(i_1 \ldots i_R)|
\end{aligned} \tag{4}$$

We will first concentrate on evaluation of the second term in Eq. (4). Let the RxN matrix $Q = [q_{ij}]$ be obtained from $Q_i = [q'_{ij}]$ and $Q_v = [q''_{ij}]$ such that $q_{ij} = q'_{ij} \cdot q''_{ij}$, $(1 \le i,j \le N)$. The following theorem gives a simple method to evaluate the second term.

Theorem 1: Let $(i_1 \ldots i_R) \in I$ and $i_{R-1} \le R < i_R$. If $k \le R$ and $k \ne i_1$, $k \ne i_2$ $\ldots$ $k \ne i_{R-1}$, then

$$|Q_i(i_1 \ldots i_R)| \cdot |Q_v(i_1 \ldots i_R)| = q_{ki_R}$$

Proof: We will prove the theorem for k=1, the proof for any other value of k follows similarly. Since Q_i and Q_v are fundamental cut-set matrices, we have $Q_i = [I_R P_i]$ and $Q_v = [I_R P_v]$ where I_R is an RxR identity matrix and P_i and P_v are RxN-R matrices. If k=1, then $Q_i(i_1 \ldots i_R) = Q_i(234 \ldots Ri_R)$ and $Q_v(i_1 \ldots i_R) = Q_v(23 \ldots Ri_R)$. We may write

$$|Q_i(23 \ldots Ri_R)| \cdot |Q_v(23 \ldots Ri_R)| = \begin{vmatrix} 0 & 0 & \cdots & 0 & q'_{1i_R} \\ 1 & 0 & \cdots & 0 & q'_{2i_R} \\ 0 & 1 & \cdots & 0 & q'_{3i_R} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & q'_{Ri_R} \end{vmatrix} \cdot \begin{vmatrix} 0 & 0 & \cdots & 0 & q''_{1i_R} \\ 1 & 0 & \cdots & 0 & q''_{2i_R} \\ 0 & 1 & \cdots & 0 & q''_{3i_R} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & q''_{Ri_R} \end{vmatrix} \tag{5}$$

By expanding the above determinants with respect to first row, we obtain $|Q_i(23 \ldots Ri_R)| \cdot |Q_v(23 \ldots Ri_R)| = (-1)^{1+R} q'_{1i_R} \cdot (-1)^{1+R} q''_{1i_R} = q'_{1i_R} q''_{1i_R}$ which is the desired result.

Theorem 2 which is a generalization of Theorem 1 will enable us to evaluate the other terms in Eq. (4). However, the significance of Theorem 2 may not be apparent until the reader has read the discussion that follows it, and possibly the example in the next section. We will concentrate on evaluation of the (k+1)th term in Eq. (4), i.e., where $(i_1 \ldots i_R) \in I$ and $i_{R-k} \le R < i_{R-k+1}$. Let $\{i_k\}$ be the set of k distinct integers less than or equal R none of which is equal to i_1 or $i_2, \ldots$ or i_{R-k}. Let P'_i and P'_v be kxN-R submatrices of P_i and P_v consisting of those rows of P_i and P_v which correspond to integers on $\{i_k\}$. By $P'_i(j_1 j_2 \ldots j_k)$, we mean the square submatrix of P'_i consisting of those columns of P'_i which correspond to branches b_{j_1} b_{j_2} $\ldots$ b_{j_k} in G_i. Similarly, we can define $P'_v(j_1 j_2 \ldots j_k)$. Note that actually the column referred to as j_r is in reality $(j_r - R)$th column of P'_i.

Theorem 2: Let $(i_1 \ldots i_R) \in I$ and $i_{R-k} \le R < i_{R-k+1}$.

$$|Q_i(i_1 \ldots i_R)| \cdot |Q_v(i_1 \ldots i_R)| = |P'_i(i_{R-k+1} \ldots i_R)| \, |P'_v(i_{R-k+1} \ldots i_R)|$$

Proof: We start by proving the theorem for the special case where $i_1 = 1$, $i_2 = 2$, $\ldots$, $i_{R-k} = R-k$, then $\{i_k\}$ contains integers R-k+1, R-k+2, ..., R. P'_i and P'_v are submatrices of P_i and P_v consisting of their last k rows. By observation similar to those mentioned in the proof of Theorem 1, we may write,

$$|Q_i(12 \ldots R\text{-}k \; i_{R-k+1} \ldots i_R)| \cdot |Q_v(12 \ldots R\text{-}k \; i_{R-k+1} \ldots i_R| =$$

$$\left|\begin{array}{c|c} I_{R-k} & P''_i \\ \hline 0 & P'_i(i_{R-k+1}\cdots i_R) \end{array}\right| \times \left|\begin{array}{c|c} I_{R-k} & P''_v \\ \hline 0 & P'_v(i_{R-k+1}\cdots i_R) \end{array}\right|$$

$$= |P'_i(i_{R-k+1}\cdots i_R)| \cdot |P'_v(i_{R-k+1}\cdots i_R)| \qquad (6)$$

which proves the theorem for the special case. To prove the theorem for the general case, one must note that if $i_1, i_2, \ldots i_R$ satisfy conditions of the hypothesis, then by changing positions of rows of $|Q_i(i_1 \ldots i_R)|$ one can obtain a determinant in the form shown in Eq. (6). Each row change causes a change in sign, but since the same changes must be made in both $Q_i(i_1 \ldots i_R)$ and $|Q_v(i_1 \ldots i_R)|$ the change in sign will be nullified. This ends the proof of Theorem 2.

Theorem 2 suggest the following general recursive procedure. To evaluate the (k+1)th term in Eq. (4), we must examine that term for all possible combinations of the integers $i_1\ i_2 \ldots i_{R-k}$ $(1 \le i_1 < i_2 \ldots < i_{R-k} \le R)$ separately. To take a particular case, let integers $(i_1\ i_2 \ldots i_{R-k})$ be $(12 \ldots R-k)$, to find all possible non-zero values of $|Q_i(12 \ldots R-k\ i_{R-k+1} \ldots i_R)| \cdot |Q_v(12 \ldots R-k\ i_{R-k+1} \ldots i_R)$ for all possible integers $i_{R-k+1}, \ldots, i_R$, $(R \le i_{R-k+1} < i_{R-k+2} \ldots < i_R \le N)$, one must find all possible non-zero values of $|P'_i(i_{R-k+1} \ldots i_R)| \cdot |P'_v(i_{R-k+1} \ldots i_R)|$. But, it can be shown that P'_i and P'_v are cut-set matrices of graphs g'_i and g'_v obtained from graphs G_i-t and G_v-t by coalescing the vertices of branches $b_1, b_2 \ldots, b_{R-k}$, respectively. By G-t, we mean complement of t in G. Hence by a row operation in each of the matrices P'_i and P'_v and by making whatever change of column position is needed in both, one obtains from P'_i and P'_v the fundamental cut-set matrices P'_{if} and P'_{vf} of g'_i and g'_v.[3] (It should be noted that any change in position of columns of one matrix, say P'_i, must be accompanied by an identical change in the position of columns in P'_v). Since the integers associated with columns of P'_{if} and P'_{vf} did not signify their position but their association with branches of g'_i and g'_v, then it is clear that for all $(j_1 \ldots j_k)$

$$|P'_i(j_1 \ldots j_k)| \cdot |P'_v(j_1 \ldots j_k)| = |P'_{if}(j_1 \ldots j_k)| \cdot |P'_{vf}(j_1 \ldots j_k)| \qquad (7)$$

However, since the k x N - R matrices P'_{if} and P'_{vf} are fundamental cut-set matrices, one can start the recursion again. The example in the next section should clarify these points, but for a better understanding the reader may want to read the second procedure for generating trees in a paper by Hakimi [9].

IV. AN EXAMPLE

Consider the active circuit with a tube and a gyrator shown in Fig. 1. Using the procedure outlined in this paper, we would like to find V_2 due to excitation V_1. The first step is to convert the excitation to a current source. (It will be shown this can be done even if $R_1 = 1/g_1 = 0$.) Fig. 2 shows the equivalent circuit of Fig. 1.

The current and the voltage graphs corresponding to the circuit of Fig. 2 is shown in Fig. 3 (a) and (b). Note that transadmittances associated with branches $b_1, b_2, \ldots, b_8$ are $y_1 = \frac{V_1 g_1}{V_2}$, $y_2 = g_1$, $y_3 = g_2$, $y_4 = sC$,

---

[3] Matrices P'_{if} or P'_{vf} should be, more accurately, called pseudo-fundamental because it may be necessary to multiply certain rows of these matrices by -1, to obtain the standard form of fundamental cut-set matrices.

$y_5 = g_p$, $y_6 = g_m$, $y_7 = 1/\alpha$, $y_8 = 1/\alpha$.

We now pick a tree t common to both graphs and write the fundamental cut-set matrices of both graphs with respect to t. A judicious choice of this tree may result in saving of time. For example, if we pick the tree $t = b_2b_3b_6b_8$, then since $b_1b_4b_5b_7$ is also a tree, we must evaluate all trees and their signs up to trees of distance 4 from t. However, if we pick the tree $t = b_2b_3b_4b_5$, we can see that every tree must either contain b_3 or b_4, or both. This is because b_3 and b_4 together form a cut-set of graph G_i of Fig. 3(a). Hence, there exists no tree of distance 4 from t, resulting in a considerable saving of computation.

The cut-set matrices Q_i and Q_v of graphs of Fig. 3(a) and (b) with respect to the tree $t = b_2b_3b_4b_5$ are

$$Q_i = \begin{array}{c} \begin{array}{cccccccc} b_2 & b_3 & b_4 & b_5 & b_1 & b_6 & b_7 & b_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \text{ and } Q_v = \begin{array}{c} \begin{array}{cccccccc} b_2 & b_3 & b_4 & b_5 & b_1 & b_6 & b_7 & b_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & -1 \end{pmatrix} \end{array} \tag{8}$$

The first term in $|Q_i Y Q_v^T|$ is due to tree, t, which is $+ y_2y_3y_4y_5$. The term due to all trees of distance one from t is evaluated, by the use of Theorem 1, and they are written as follows

$$y_2y_3y_4y_5 \quad \frac{1}{y_2}(0) + \frac{1}{y_3}(-y_7+y_8) + \frac{1}{y_4}(-y_7+y_8) + \frac{1}{y_5}(y_6) \,. \tag{9}$$

The above term is due to all trees containing exactly 3 out of 4 branches of t. The term due to all trees containing 2 out of 4 branches of t is evaluated by Theorem 2. We first find all terms due to trees containing b_2 and b_3. To do this, we pick submatrices P'_i and P'_v consisting of the last 2 rows of P_i and P_v. These submatrices are

$$P'_i = \begin{array}{c} \begin{array}{cccc} b_1 & b_6 & b_7 & b_8 \end{array} \\ \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \text{ and } P'_v = \begin{array}{c} \begin{array}{cccc} b_1 & b_6 & b_7 & b_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & -1 \end{pmatrix} \end{array} \tag{10}$$

By rearranging the columns and performing some row operations, the matrices of Eq. (10) become

$$P''_i = \begin{array}{c} \begin{array}{cccc} b_7 & b_6 & b_8 & b_1 \end{array} \\ \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \text{ and } P''_v = \begin{array}{c} \begin{array}{cccc} b_7 & b_6 & b_8 & b_1 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \end{pmatrix} \end{array} \tag{11}$$

We multiply the first row of P''_i by (-1), to get both matrices in the form of fundamental cut-set matrices.

$$P'''_i = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \text{ and } P'''_v = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \end{pmatrix} \tag{12}$$

By Theorem 1 and Eq. (4), the sum of tree admittance product of all trees that are in two graphs whose cut-set matrices are P'''_i and P'''_v is

$$y_7y_6 + y_7y_6 \left[\frac{1}{y_7}(-y_8) + \frac{1}{y_6}(-y_8) \right] \tag{13}$$

Multiplying the sum of Eq. (13) by (-1) y_2y_3, (the (-1) is due to the change in sign brought about in going from Eq. (11) to Eq. (12), we obtain the sum of tree admittance products all of which contain y_2y_3

$$-y_2y_3y_7y_6 + y_2y_3y_6y_8 + y_2y_3y_7y_8 \tag{14}$$

Similarly the sum of tree admittance products all of which contain y_2y_4, y_2y_5, y_3y_4, y_3y_5, and y_4y_5 are computed.

$$-y_2y_4y_6y_7 + y_2y_4y_7y_8 + y_3y_4y_7y_8 + y_3y_5y_1y_7 + y_3y_5y_7y_8 + y_4y_5y_1y_7 + y_4y_5y_7y_8 \tag{15}$$

There is no term containing y_2y_5, because the second and the third rows are of Q_i are linearly dependent. For identical reason there are no terms containing y_2 (but not $y_3y_4y_5$) and y_5 (but not $y_2y_3y_4$). The terms containing y_3 and y_4 are

$$y_3y_1y_7y_6 + y_3y_7y_6y_8 - y_3y_1y_7y_8 + y_4y_1y_7y_6 - y_4y_1y_7y_8 \tag{16}$$

Since there are no trees at a distance of four from t, there are no other terms. Due to Eq. (2), the sum of all admittance products obtained in Eqs. (9), (13), (14) and (15) must be equal to zero. From that equation y_1 is calculated to be

$$y_1 = \frac{\begin{array}{c} y_2y_3y_4y_5+y_2y_4y_5(-y_7+y_8)+y_2y_3y_5(-y_7+y_8)+y_2y_3y_4y_6+y_2y_3(-y_7y_6+y_6y_8+y_7y_8) \\ +y_2y_4(-y_7y_6+y_7y_8)+y_3y_4y_7y_8+y_3y_5y_7y_8+y_4y_5y_7y_8+y_3y_7y_8y_6 \end{array}}{-y_3y_5y_7-y_4y_5y_7-y_3y_6y_7+y_3y_7y_8-y_4y_6y_7+y_4y_7y_8} \tag{17}$$

Substituting for $y_1, y_2, \ldots, y_8$ their values, we obtain

$$\frac{V_1g_1}{V_2} = \frac{\begin{array}{c} sC(g_1g_2g_m+g_1g_2g_p-g_1g_m\frac{1}{\alpha} + g_1\frac{1}{\alpha^2} + g_2\frac{1}{\alpha^2} + g_p\frac{1}{\alpha^2}) \\ +g_1g_2\frac{1}{\alpha^2} +g_2g_p\frac{1}{\alpha^2} +g_2g_m\frac{1}{\alpha^2} \end{array}}{sC(\frac{1}{\alpha^2} - g_p\frac{1}{\alpha} - g_m\frac{1}{\alpha})-g_2g_p\frac{1}{\alpha} - g_2g_m\frac{1}{\alpha} + g_2\frac{1}{\alpha^2}} \tag{18}$$

From Eq. (18), $\frac{V_2}{V_1}$, the open circuit voltage transfer function can be found. If one takes the limit as $g_1 \rightarrow \infty$, one obtained the open circuit voltage transfer function when $R_1 = 0$

$$\left.\frac{V_2}{V_1}\right|_{R_1 = 0} = \frac{sC(\frac{1}{\alpha^2} - g_p\frac{1}{\alpha} - g_m\frac{1}{\alpha}) - g_2g_p\frac{1}{\alpha} - g_2g_m\frac{1}{\alpha} + g_2\frac{1}{\alpha^2}}{sC(g_2g_m+g_2g_p-g_m\frac{1}{\alpha} + \frac{1}{\alpha^2}) + g_2\frac{1}{\alpha^2}} \tag{19}$$

It is true that the number of computer operations is very large, but other available general methods have the same disadvantage. However, in this procedure, the memory space needed is relatively small because only the original matrices must remain in the memory permanently.

V. CONCLUSIONS, FURTHER PROBLEMS AND REMARKS

To our judgement, the method presented here is the best general topological method available. We have purposely avoided deriving topological formulas, because they are known and they do not further facilitate the problem of analysis of active networks. The method presented here can also be used in connection with Talbot's method, but we prefer basing it on Percival-Mayeda's formulation, because here there is no need of keeping track of trees that must be paired. Our method, as well as Mayeda-Seshu's method, suffer from one difficulty, and that is, in a worst case, one may have to consider 2^R possible combinations of branches of t separately. This is a large number. This problem can be partially rectified by the following considerations:

(1) Proper choice of starting tree t can be of significant help. For example, an even better starting tree for the problem considered in Section 4 would be $t = b_3b_4b_7b_8$. Since graph G_i-t has rank two, there would be no trees of distance larger than two from t. Thus an optimum choice of starting tree would be a tree t such that rank of G-t is minimum. This problem remains unsolved. The problem at hand is even more difficult for the optimum choice of the starting tree depends on both G_i and G_v. But it is very easy to pick a tree that is a great improvement over a random choice of starting tree.

(2) In our procedure for generating trees, we first find all trees that contain all but one (any one) branch of starting tree t, then all trees that contain all but 2 branches of tree t and so forth. Let tree t consist of branches $b_1, b_2, \ldots b_R$. In generating all trees that contain all but k branches of t, we eliminate a particular k branches of t, say $b_{i_1}, b_{i_2}, \ldots, b_{i_k}$ and try to construct new trees in which these k branches are replaced by k branches from G-t. But suppose there does not exist any such trees. (Such a situation arises only if branches $b_{i_1}, b_{i_2}, \ldots, b_{i_k}$ contain a cut-set). Then there certainly does not exist any trees that do not contain any of the branches in $b_{i_1}, b_{i_2}, \ldots b_{i_k}, b_{i_{k+1}}$ of t. However in our procedure we do not take advantage of this knowledge. This idea may be abstractly formulated as the following significant problem: Consider a set S of n elements. These elements may be identified by integers $1, 2, \ldots, n$. We want to generate all possible sets which contain $\leq$ k elements of S and each satisfy decision a rule U. In addition, we know that if set s consisting of elements of S does not satisfy rule U, then any set that contains s does not satisfy U. We want to avoid generating unacceptable sets as much as possible for application of decision rule U is time consuming. It can be shown that the following procedure generates all of the desired subsets (without duplication and the use of memory) and substantially reduces the number of times the decision rule U must be applied.

We start by picking sets 1,12,123 ... . We continue as follows: (a) If the set $i_1 i_2 \ldots i_p$, ($i_p < n$ and $p < k$) satisfies rule U, then the next set is: $i_1 i_2 \ldots i_p (i_p+1)$. (b) If the set $i_1 i_2 \ldots i_p$ ($i_p < n$ and $p \leq k$) does not satisfy rule U, then the next set is: $i_1 i_2 \ldots i_{p-1} (i_p+1)$, (c) If the set $i_1 i_2 \ldots i_p$, ($i_p = n$, and $p \leq k$)then the next set is: $i_1 i_2 \ldots i_{p-2} (i_{p-1}+1)$, and (d) If in the set $i_1 i_2 \ldots i_p$, ($i_p < n$ and $p = k$,)then the next set is: $i_1 i_2 \ldots i_{p-1} (i_p+1)$. This algorithm can be improved and simplied if the use of memory is allowed.

Although these two considerations are of some value, it seems that a "partitioning technique" must be used for circuits consisting of more than twenty nodes. No acceptable partitioning technique for topological analysis of active circuits is as yet available.

If one is interested in obtaining a network function as a function of only complex variable s but not explicit function of circuit parameters (i.e., parameter values are given as numbers), it seems that the best available method is the state space technique describing by Pottle [19]. However, in many applications such as design of filters and amplifiers, where one picks the configurations and uses a non-linear programming technique to obtain the parameter values [20,21], one needs the network function explicitly as a function of s and the circuit parameters . In such a situation topological methods still seem best.

REFERENCES

[1] W.S. Percival, "Solution of Passive Electrical Networks by Means of Mathematical Trees", *T. Inst. Elec. Engrs*., (London) 100, part III, pp. 143-150, 1953.

[2] W.S. Percival, "Graphs of Active Networks", *Proc* of *IRE*, 102, part C, pp. 270-278, 1955.

[3] W. Mayeda and S. Seshu, "Topological Formulas for Network Functions", Bulletin 446 of the Engineering Experiment Station, Univ. of Ill., Urbana, Ill., 1956.

[4] G. Kirchhoff, *Poggendorf Ann. Physik*, 72, pp. 497-508, 1847; English Translation, IRE Trans., CT-5, pp. 4-7, 1958.

[5] J.C. Maxwell, "*Electricity and Magnetism*, 1", Chapter 4 and Appendix, pp. 403-410. Clarendon Press, 1892.

[6] C.L. Coates, "General Topological Formulas for Linear Network Functions", *Trans. IRE*, CT-5, pp. 30-42, 1958.

[7] W. Mayeda, "Topological Formulas for Non-reciprocal Networks and Networks with Transformers", *Proc*. of *NEC*, 14, pp. 631-643, 1958.

[8] W. Mayeda, "Reducing Computation Time in the Analysis of Networks by Digital Computers", *Trans. IRE*, CT-6, pp. 136-137, 1959.

[9] S.L. Hakimi, "On Trees of a Graph and Their Generation", *J. Franklin Inst*., 270, pp. 347-359, 1961.

[10] S.L. Hakimi and D.G. Green, "Generation and Realization of Trees and k-Trees", *Trans. IEEE*, CT-11, pp. 247-255, 1964.

[11] A. Talbot, "Topological Analysis of General Linear Networks", *Proc of IEEE Trans*., CT-12, pp. 170-180, 1965.

[12] W. Mayeda and S. Seshu, "Generation of Trees without Duplication", *IEEE Trans*., CT-12, pp. 181-185, 1965.

[13] D.P. Brown, "New Topological Formulas ...", to be published in the IEEE Trans. on Circuit Theory.

[14] A.A. Ali, "On the Sign of a Tree Pair", *IEEE Trans*. CT-11, pp. 294-296; 1964.

[15] I.T. Frisch and W.H. Kim, "Properties of 2-semi-isomorphic Graphs and Their Applications: Active Network Analysis", *J. Math. Phys*., Vol. 2, pp. 627-635; Aug. 1961.

[16] W.K. Chen, "Toplogical Analysis of Active Networks", *IEEE Trans*., CT-12, pp. 85-91; 1965.

[17] O. Wing, "Enumeration of Trees", *IEEE Trans*., CT-10, pp. 127-128; 1963.

[18] S. Seshu and M.B. Reed, "*Linear Graphs and Electrical Networks*", Addison-Wesley, Reading, Mass., 1961.

[19] C. Pottle, "Comprehensive Active Network Analysis by Digital Computer-a State-Space Approach", *Proc of 3rd Allerton Conference*, Univ. of Ill., pp. 659-668, Oct. 1965.

[20] L.S. Lasdon and A.D. Warren, "Optimal Design of Filters with Bounded and Lossy Elements", *IEEE Trans*., CT-13, pp. 175-187, June 1966.

[21] P.O. Scheibe and E.A. Huber, "The Application of Carrol's Optimization to Network Synthesis", *Proc of 3rd Allerton Conference*, Univ. of Ill., pp. 182-191, Oct. 1965.

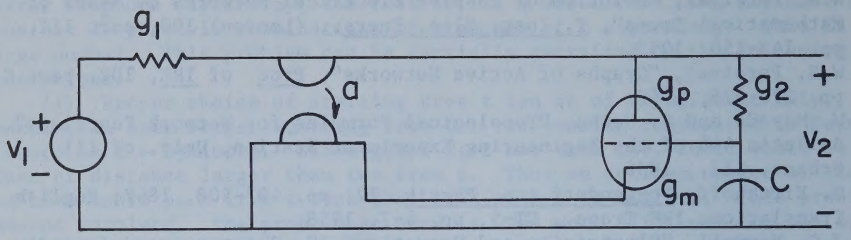

g1
a
+
v1
−
gp
g2
+
v2
gm
C
−

Fig. 1

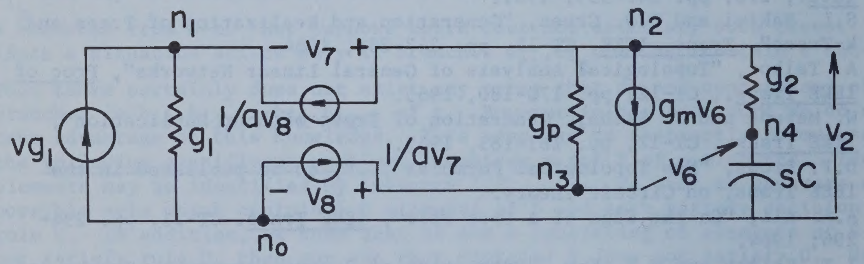
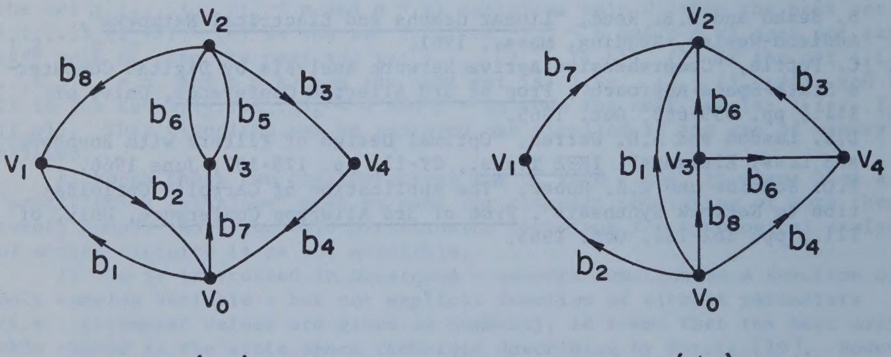

n1
− v7 +
n2
vg1
g1
1/av8
gp
gmv6
g2
n4
v2
1/av7
n3
v6
sC
− v8 +
n0

Fig. 2


v2
b8
b6
b5
b3
v1
v3
v4
b2
b7
b4
b1
v0
v2
b7
b3
b6
v1
b1
v3
v4
b6
b8
b4
b2
v0

(a) (b)

Fig 3 - (a) Current graph and (b) Voltage graph corresponding to the circuit of Fig. 2

ON THE GENERATION OF NON-SINGULAR SUBMATRICES AND THEIR CORRESPONDING SUBGRAPHS*

WAI-KAI CHEN
Department of Electrical Engineering
Ohio University
Athens, Ohio

ABSTRACT

The intimate relationship between subgraphs of a graph and the maximal non-singular submatrices of the associated matrix of the graph is well-known. In this paper, a procedure for enumerating all the maximal non-singular submatrices of a given matrix without duplications is presented. The procedure is then applied to the generation of all possible trees, two-trees, complete trees, complete two-trees, directed trees, and directed two-trees without duplications. The method can easily be programmed for a digital computer, and it will reduce the computing time necessary to generate all these subgraphs more than any of the technique available to date.

1. INTRODUCTION

The usefulness of topological formulae in the analysis of a complex electrical network depends largely upon the efficiency with which the trees of the corresponding graph are generated. The main difficulties lie in the fact that one has to find a method which requires that the number of operations increases linearly with the number of vertices and not exponently, and which does not need to check every new tree against all the previously found trees in order to eliminate duplications, because the exhausting procedure is an extremely time-consuming process as the number of trees in practical networks may be of the order of a million. Several methods for generating trees of a graph are available,[4-6, 11-16] but none of these procedures is completely satisfactory. They all suffer either one or two of the disadvantages mentioned above. Recently, a procedure based on the fundamental cutset matrix for generating trees is given by Mayeda and Seshu,[1] which shows considerable improvement over the older ones.[4-6, 11-16]

The intimate relationship between trees of a graph and the maximal non-singular submatrices of the corresponding cutset matrix is well-known, and needs no further elaboration. The purpose of this paper is to describe a procedure for enumerating all the maximal non-singular submatrices of a given matrix without duplications, and then apply these results for generating all possible trees, complete two-trees, directed trees and directed two-trees without duplications. The method presented here can be programmed easily for a digital computer, and it will reduce the computing time necessary to generate all trees more than any of the techniques available to date. This method may be considered as a simplified version of Mayeda and Seshu's technique,[1] and indeed it was their paper which suggested the present investigation.

The work reported in this paper was supported jointly by the National Science Foundation under the Grant No. GK-1095 and Ohio University.

2. THE ENUMERATION OF MAJOR SUBMATRICES OF A MATRIX WITHOUT DUPLICATIONS

Let F be an arbitrary matrix of order m by n and of rank m. Assume $m \leq n$. Otherwise we may consider the transpose of F instead of F. A submatrix F_s is said to a major submatrix of F if F_s is a non-singular submatrix of order m in F. In other words, the determinant of F_s is a non-vanishing major determinant of F. In order that the columns of F which will form a major submatrix be expressed explicitly, the symbol $F(j_1 j_2 \cdots j_m)$ will be used to denote the major submatrix of F consisting of the columns $j_1, j_2, \ldots,$ and j_m of F. For convenience, we refer to a matrix containing a diagonal major submatrix as a matrix in normal form. In the following, we assume all the given matrices are in normal form. This assumption is not a serious restriction, as it can easily be shown that to every matrix F of order m by n $(m \leq n)$ and of rank m, there corresponds a matrix in normal form which can be obtained from F by using only the following row operations:

(a) Addition to one row of a multiple of another row, and
(b) permutation of rows.

Since the matrix thus obtained generates the same linear vector space as F does, they would have the same set of non-vanishing major determinants with at most a permutation of rows. However, the relative signs of the major determinants are preserved.

For a given matrix F in normal form, $F_d(k_1 k_2 \ldots k_m)$ is used to represent the defining diagonal major submatrix $F(k_1 k_2 \ldots k_m)$ of F; and $\overline{F}(k_1 k_2 \ldots k_m)$ the matrix obtained from F by deleting all the columns of F corresponding to the columns $k_1, k_2, \ldots,$ and k_m; and $F(i_1 i_2 \ldots i_q; j_1 j_2 \cdots j_q)$ the submatrix of F consisting of those rows of F corresponding to the rows $i_1, i_2, \ldots,$ and i_q, and those columns of F corresponding to the columns $j_1, j_2, \ldots, j_q$. For convenience, the columns of the given matrix F are assumed to have been labelled by one of the integers from 1 to n, and the rows of F are labelled in such a way that the ith row of F is denoted by k_i for $i = 1, 2, \ldots, m$ where k_i are the columns in $F_d(k_1 k_2 \ldots k_m)$. Throughout this paper the term matrix means a matrix which has been labelled according to rules just outlined. Also, all the integers in the parentheses are assumed to appear in their natural order. Let

$$i_1 i_2 \cdots i_p \frac{j_1 j_2 \cdots j_q}{i_1 i_2 \cdots i_q} = \begin{cases} i_{q+1} i_{q+2} \cdots i_p j_1 j_2 \cdots j_q & \text{for } q \geq 1 \\ i_1 i_2 \cdots i_p & \text{for } q = 0 \end{cases}$$

where $0 \leq q \leq p$. The following example will illustrate this point:

$$F(156 \tfrac{28}{16}) = F(258).$$

Notice that the integers in the parentheses are again assumed to appear in their natural order. With these definitions, it follows that

$$F(k_1 k_2 \cdots k_m \frac{j_1 j_2 \cdots j_q}{i_1 i_2 \cdots i_q}), \quad 1 \leq q \leq m$$

is a major submatrix of F other than $F_d(k_1k_2\ldots k_m)$ if, and only if, $F(i_1i_2\cdots i_q; i_1i_2\cdots i_q)$ is a non-singular submatrix of $\overline{F_d(k_1k_2\ldots k_m)}$. Consequently, the following theorem is obvious.

Theorem 1: Let F be a matrix of order m by n ($m \leqslant n$) and of rank m in normal form. Then the set M of all major submatrices of F is given by

$$M = \bigcup_{q=0}^{r} \bigcup_{i_1i_2\cdots i_q} F(k_1k_2\ldots k_m \frac{i_1i_2\cdots i_q}{i_1i_2\cdots i_q})$$

where $\overline{F(i_1i_2\ldots i_q; i_1i_2\cdots i_q)}$ is a non-singular submatrix of $\overline{F_d(k_1k_2\ldots k_m)}$; r is the rank of $\overline{F_d(k_1k_2\ldots k_m)}$; and the second union is taken over all possible combinations of the form $i_1i_2\ldots i_q$.

The problem of enumerating all major submatrices of F is now reduced to the problem of enumerating all the non-singular submatrices of orders 1,2,..., and r in $\overline{F_d(k_1k_2\ldots k_m)}$. The non-singular submatrices of order 1 are very easy to find; they are the non-zero elements in $\overline{F_d(k_1k_2\ldots k_m)}$. For the non-singular submatrices of order greater than 1, the whole process may be repeated. For example, consider the submatrix

$$F(i_1i_2\cdots i_q; i_1i_2\cdots i_{n-m})$$

where $2 \leqslant q \leqslant r$, and $i_1, i_2, \ldots, i_{n-m}$ and $k_1, k_2, \ldots, k_m$ are complementary indices of the integers 1, 2, ..., n. This matrix is first put in normal form, and by Theorem 1 all the major submatrices of this matrix can be obtained. If all the submatrices of $\overline{F_d(k_1k_2\ldots k_m)}$ of the above type are considered, the totality of all these major submatrices is the set of all non-singular submatrices of order q in $\overline{F_d(k_1k_2\ldots k_m)}$. Note that in each of the above steps, the same process may again be repeated until all the non-singular submatrices can be obtained by inspection or are of order 1.

3. THE ENUMERATION OF TREES AND COTREES* OF A GRAPH WITHOUT DUPLICATIONS

As mentioned earlier, there exists a one-to-one correspondence between trees (cotrees) of a graph and the major submatrices of the corresponding cutset (circuit) matrix of the graph. By using these relationships and Theorem 1, a method of generating trees (cotrees) without duplications can now be stated. However, before doing this, the graph and the associated matrices are assumed to have been labelled properly. A labelled graph is a graph in which all the edges are labelled by one of the integers from 1 to e such that the edge labelled k is associated with the kth column of the cutset matrix or circuit matrix of the graph where e is the number of edges of the graph, and also all the nodes are labelled by $n_1, n_2, \ldots, n_v$ where v is the number of nodes of the graph. Unless otherwise specified, throughout this paper the term graph means a labelled graph.

Theorem 2: If Q is a cutset matrix in normal form of a connected graph G of v nodes and e edges, with v-1 rows and rank v-1, then the set T of all possible distinct trees of G is given by

$$T = \left\{ k_1k_2\ldots k_{v-1} \frac{i_1i_2\cdots i_q}{i_1i_2\cdots i_q}; \quad \text{for } q = 0, 1, \ldots, r \right\}$$

---

*It is also called the chord-sets.

where $Q(i_1i_2\dots i_q;\ j_1j_2\dots j_q)$ is a non-singular submatrix of $\overline{Q_d(k_1k_2\dots k_{v-1})}$; and r is the rank of $\overline{Q_d(k_1k_2\dots k_{v-1})}$.

Since only the columns of the major submatrices of Q are needed, not the value of the determinants of the major submatrices, we may interchange the columns of Q before applying the algorithm; i.e., we may assume that the diagonal submatrix of Q appears in the first v-1 columns of A or in the last v-1 columns of Q. However, if the algorithm is used to generate all the major submatrices of a cutset matrix of the corresponding graph associated with a given network with active and mutual elements, then only the row operations defined in Section 2 are allowed. Since these submatrices are in one-to-one correspondence with the complete trees of the associate graphs of Coates[2] and Mayeda,[3] and also the relative signs of the determinants of the major submatrices are preserved in both the matrix and the matrix in normal form. (Only row operations are allowed.) the algorithm may readily be used in generating all complete trees.

Recently, Fujisawa[4] presented a method of obtaining all possible cotrees and the sign of the cotrees of the corresponding graph associated with the loop-impedance matrix of a given active network, dual to the Coates-Mayeda formulae. However, his method is still too complicated and is not easily adaptable to the digital computer. The method described above may be used, with minor modification, for this purpose, and is stated as

Theorem 3: If B is a circuit matrix in normal form of a connected graph G of v nodes and e edges, with e-v+ 1 rows and rank e-v+ 1, then the set C of all possible distinct cotrees of G is given by

$$C = \left\{ k_1k_2\dots k_{e-v+1}\frac{i_1i_2\cdots i_q}{i_1i_2\cdots i_q}\ ; \qquad \text{for } q = 0, 1, \dots, r \right\}$$

where $B(i_1i_2\dots i_q;\ j_1j_2\dots j_q)$ is a non-singular submatrix of $\overline{B_d(k_1k_2\dots k_{e-v+1})}$; and r is the rank of $\overline{B_d(k_1k_2\dots k_{e-v+1})}$.

In the following a definition similar to the one defined by Watanabe[5] for trees of a graph is needed. For a subgraph g of G, the symbol N(g) is used to denote the number of edges of g. If g_1 and g_2 are two subgraphs of G with $N(g_1) = N(g_2)$, then the distance between the subgraphs g_1 and g_2, denoted by $d(g_1,g_2)$, is defined as the number of edges appeared in one but not in the other, i.e., $d(g_1,g_2)= \frac{1}{2}N(g_1\oplus g_2)$ where $g_1\oplus g_2$ is the ring sum of g_1 and g_2. In particular, g_1 and g_2 could be trees or cotrees of a graph.

Theorem 4: If Q is a cutset matrix in normal form of a connected graph G of v nodes and e edges, with v-1 rows and rank v-1, then the set T_q of all possible distinct trees of distance q from the tree t_o in G is given by

$$T_q = \left\{ k_1k_2\dots k_{v-1}\frac{i_1i_2\cdots i_q}{i_1i_2\cdots i_q} \right\}$$

where $q \lesssim r$; $Q(i_1i_2\dots i_q;\ j_1j_2\dots j_q)$ is a non-singular submatrix of $\overline{Q_d(k_1k_2\dots k_{v-1})}$; r is the rank of $\overline{Q_d(k_1k_2\dots k_{v-1})}$; and $t_o = k_1k_2\dots k_{v-1}$.

Corollary 1: The number of trees of distance q from t_o in G is equal to the number of non-singular submatrices of order q in $\overline{Q_d(k_1k_2\dots k_{v-1})}$.

Corollary 2: The number of trees of distance 1 from t_o in G is equal to the number of non-zero entries in $\overline{Q_d(k_1k_2\dots k_{v-1})}$.

Although the dual results of the above discussion are only trivially different from those given above, for completeness, they are stated as follows:

Theorem 5: If B is a circuit matrix in normal form of a connected graph G of v nodes and e edges, with e-v+ 1 rows and rank e-v+ 1, the set C_q of all possible distinct cotrees of distance q from the cotree c_o in G is given by

$$C_q = \left\{ k_1k_2\ldots k_{e-v+1} \frac{i_1i_2\ldots i_q}{i_1i_2\ldots i_q} \right\}$$

where $q \leq r$; $B(i_1i_2\ldots i_q; i_1i_2\ldots i_q)$ is a non-singular submatrix of $\overline{B_d(k_1k_2\ldots k_{e-v+1})}$; r is the rank of $\overline{B_d(k_1k_2\ldots k_{e-v+1})}$; and $c_o = k_1k_2\ldots k_{e-v+1}$.

Corollary 3: The number of cotrees of distance q from c_o in G is equal to the number of non-singular submatrices of order q in $\overline{B_d(k_1k_2\ldots k_{e-v+1})}$.

Corollary 4: The number of cotrees of distance 1 from c_o in G is equal to the number of non-zero entries in $\overline{B_d(k_1k_2\ldots k_{e-v+1})}$.

It is interesting to note that the technique presented here can equally well be applied to both Q and B of G. However, the dual procedure based on B of the method of Mayeda and Seshu[1] is by no means trivial. Since, for a connected graph G of nullity > 0, a cotree is the unique complement of a tree in G, it follows that two trees are of distance q if and only if the corresponding two cotrees are of distance q. The following corollaries are a direct consequence of the above discussion.

Corollary 5: Let $c_o = h_1h_2\ldots h_{e-v+1}$ be the corresponding cotree of the tree $t_o = k_1k_2\ldots k_{v-1}$ in G. Then the number of non-singular submatrices of order q in $\overline{Q_d(k_1k_2\ldots k_{v-1})}$ is equal to the number of non-singular submatrices of order q in $\overline{B_d(h_1h_2\ldots h_{e-v+1})}$. In other words, the number of trees (cotrees) of distance q from t_o (c_o) is equal to the number of non-singular submatrices of order q in $\overline{B_d(h_1h_2\ldots h_{e-v+1})}$ or in $\overline{Q_d(k_1k_2\ldots k_{v-1})}$ where $h_1, h_2, \ldots, h_{e-v+1}$ and $k_1, k_2, \ldots, k_{v-1}$ are the complementary sets of the integers $1, 2, \ldots, e$.

Corollary 6: The subgraph corresponding to the columns $k_{i_1}, k_{i_2}, \ldots, k_{i_t}$ of $\overline{Q_d(k_1k_2\ldots k_{v-1})}$ can be made part of a cortee if and only if the rank of the matrix obtained by taking the rows $k_{i_1}, k_{i_2}, \ldots, k_{i_t}$ of $\overline{Q_d(k_1k_2\ldots k_{v-1})}$ is equal to t where $t \leq v-1$.

Since a subgraph can be made part of a tree if and only if it contains no circuits, the dual result may now be stated in a slightly different form.

Corollary 7: (due to Hakimi) The subgraph corresponding to the columns $k_{i_1}, k_{i_2}, \ldots k_{i_t}$ of $\overline{B_d(k_1k_2\ldots k_{e-v+1})}$ is circuitless if and only if the rank of the matrix obtained by taking the rows $k_{i_1}, k_{i_2}, \ldots, k_{i_t}$ of $\overline{B_d(k_1k_2\ldots k_{e-v+1})}$ is equal to t where $t \leq e-v+1$.

Note that in all of the above discussion it is assumed that Q and B are in normal form. However, it is not necessary they be a fundamental cutset matrix or a fundamental circuit matrix of the graph. If they are, they are already in normal form, but the converse is not in general true.

Corollary 5 may be considered as a generalized theorem of a recent work of Hakimi,[6] i.e., it reduces to a theorem of Hakimi when $q = 1$ and B is the fundamental circuit matrix with respect to the tree t_o of the graph G.

In the light of Theorem 4, it is easy to see that the set T_p of all possible distinct trees of distance p from the reference tree t_o is, in terms of the notation of Mayeda and Seshu,[1] given by

$$T_p = \bigcup_{i_1i_2\ldots i_p} T^{e_{i_1}e_{i_2}\ldots e_{i_p}}$$

where p is a positive integer. Again, the elements of T_p are in one-to-one correspondence with the non-singular submatrices of order p in $\overline{Q_d(k_1k_2\ldots k_{v-1})}$

where $t_o = k_1k_2 \ldots k_{v-1}$. In fact, the above equation is the main result of a recent work of Meyeda and Seshu.[1]

Earlier, we refer to a matrix F containing a diagonal major submatrix as a matrix in normal form, and assume all the given matrices are in normal form. Actually, this restriction can be relaxed somewhat. For convenience, a matrix F containg a submatrix F_p is said to be a matrix in proper form if F_p is a major submatrix of F and such that each column of F_p contains exactly one non-zero element, i.e., F_p can be put in diagonal form by merely permutating the rows of F_p. This means that to every matrix F of order m by n ($m \leqslant n$) and of rank m, there corresponds a matrix in proper form which can be obtained from F by using only the operations of adding to one row of a multiple of another row. Consequently, the value of the major determinants of F and the value of the corresponding major determinants of F in proper form are equal. It is not difficult to see that all the theorems and corollaries given so far are also true if the term "in normal form" is being replaced by the term "in proper form" provided the rows of a given matrix F are labelled properly, i.e., if $F_p = F(k_1k_2 \ldots k_m)$, then the jth row of F is labelled by k_i if the non-zero entry of column k_i appears at the jth row. This result is useful if the technique presented here is used to generate all complete trees and complete 2-trees of a network. Since no significant result is added, further elaboration is unnecessary.

4. THE ENUMERATION OF TWO-TREES OF A GRAPH WITHOUT DUPLICATIONS

The procedure of enumerating trees of a graph without duplications can now be extended to the generation of two-trees of a graph without duplications. Very often, two-trees in which certain designated nodes are required to be in different components are used. For example, $t_{ab,c}$ is the symbol for a two-tree in which the nodes a and b are in one component, and the node c in one component. For a given (labelled) graph G of v nodes and e edges, let A be the complete incidence matrix of G in which all the nodes of the graph G are represented, i.e., A is of order v by e. Let $A_{n_in_jn_k}$ be the matrix obtained from A by the deleting of the rows of A corresponding to the nodes n_i, n_j, and n_k of G. If $Z = \left[z_{ij}\right]_{t \times u}$ and $W = \left[w_{ij}\right]_{t \times u}$ are two matrices of integers mod 2, then the intersection of Z and W, denote by $Z \cap W$, is the matrix of order t by u such that the ijth element of $Z \cap W$ is equal to $z_{ij}w_{ij}$. With these definitions, the following theorem can now be stated.

Theorem 6: Let A be the complete incidence matrix of a given graph G of v nodes and e edges, and also let Z and W be used to denote the corresponding matrices $A_{n_in_k}$ and $A_{n_jn_k}$ in normal form, respectively, such that their defining diagonal major submatrices consist of the same columns of A. If $Z(i_1i_2 \ldots i_q; j_1j_2 \ldots j_q)$ and $W(i_1i_2 \ldots i_q; j_1j_2 \ldots j_q)$ are the corresponding non-singular submatrices of order q in $\overline{Z_d(k_1k_2 \ldots k_{v-2})}$, and $\overline{W_d(k_1k_2 \ldots k_{v-2})}$, respectively, then the set $T^q_{n_in_j,n_k}$ of all distinct two-trees of the form $t_{n_in_j,n_k}$ of distance q from the reference two-tree $t^o_{n_in_j,n_k} = k_1k_2 \ldots k_{v-2}$ is given by

$$T^q_{n_in_j,n_k} = \left\{ k_1k_2 \ldots k_{v-2} \frac{i_1i_2 \ldots i_q}{j_1j_2 \ldots j_q} \right\}, \text{ for } q = 0, 1, \ldots, r$$

where r is the minimum rank of either $\overline{Z_d(k_1k_2 \ldots k_{v-2})}$ or $\overline{W_d(k_1k_2 \ldots k_{v-2})}$.

Proof: Since two-trees $t_{n_in_j,n_k}$ of G are in one-to-one correspondence with the non-zero products of the corresponding major determinants of $A_{n_in_k}$ and the transpose of $A_{n_jn_k}$, the theorem follows directly from the fact that the determinant of a matrix is equal to the determinant of the transpose of the matrix. This completes the proof of the theorem.

Corollary 8: Let $Y = Z \cap W$. Then

$$T^1_{n_i n_j, n_k} = \left\{ k_1 k_2 \cdots k_{v-2} \frac{j}{i} \right\}$$

where $Y(i; j)$ is a non-singular submatrix of order 1 (non-zero entry) in $\overline{Y_p(k_1 k_2 \cdots k_{v-2})}$.

5. THE ENUMERATION OF COMPLETE TREES AND COMPLETE TWO-TREES OF A NETWORK WITHOUT DUPLICATIONS

Coates[2] and Mayeda[3] have shown that the network functions of an active electrical network with mutual inductances can be evaluated by finding the sets of all complete trees and complete two-trees of the associated current- and voltage- graphs. The technique presented earlier can again be used for generating such subgraphs without duplications.

Earlier, we refer to a matrix F containing a major submatrix F_p such that each column of F_p contains exactly one non-zero element as a matrix in proper form. In order to express the columns explicitly, the symbol $F_p(k_1 k_2 \cdots k_m)$ is used to represent such defining major submatrix where m is the number of rows of F. As mentioned earlier, to every matrix, there corresponds a matrix in proper form which can be obtained by using only the operations of adding to one row of a multiple of another row. We shall use this fact in determining the sign of complete trees and two-trees.

The intersection of two matrices H and R of the same order of the integers 1, 0, and -1 is defined exactly the same as in the previous section except it is now over the real field.

Let A^i and A^v be the complete incidence matrices of the associated current graph and voltage graph, respectively, of v nodes and e edges, of a given network such that the jth column of both A^i and A^v represents the jth edge of the graphs. Also let $A^{i*}_{n_r}$ and $A^{v*}_{n_r}$ be used to denote the matrices $A^i_{n_r}$ and $A^v_{n_r}$ in proper form, respectively, such that their defining major submatrices (corresponding to F_p in the above discussion) consist of the same columns where n_r is the reference node of the graph, and as usual the subscripts indicate that the rows corresponding to the subscripts have been deleted. Assume H is the matrix obtained from $A^{v*}_{n_r}$ by row interchanges and multiplying rows by (-1) if necessary such that H and $A^{i*}_{n_r}$ have the same defining major submatrices. The number of such row interchanges plus the number of (-1)'s needed is represented by N_1.

For convience, if a complete tree (or two-tree) has an associated minus sign, a "-" is prefixed for the term. Otherwise, it represents a complete tree (or two-tree) with plus sign. Note that the associated graph G of a given network is labelled the same as before except that the corresponding current element and voltage element are labelled by the same integer. With these definitions, the following results can now be stated.

Theorem 7: With the same notation used above, the corresponding non-singular submatrices of $\overline{R_p(k_1 k_2 \cdots k_{v-1})}$ and $\overline{H_p(k_1 k_2 \cdots k_{v-1})}$, respectively, are in one-to-one correspondence with the complete trees of G where $R = A^{i*}_{n_r}$. In other words, if $R(i_1 i_2 \cdots i_q; j_1 j_2 \cdots j_q)$ and $H(i_1 i_2 \cdots i_q; j_1 j_2 \cdots j_q)$ are the non-singular submatrices of $\overline{R_p(k_1 k_2 \cdots k_{v-1})}$ and $\overline{H_p(k_1 k_2 \cdots k_{v-1})}$, respectively, then the set T_q of all distinct complete trees of G of distance q from the reference tree $t_o = k_1 k_2 \cdots k_{v-1}$ is given by

$$T_q = \left\{ (-1)^{N_1 + N_x} k_1 k_2 \cdots k_{v-1} \frac{j_1 j_2 \cdots j_q}{i_1 i_2 \cdots i_q} \right\}, \quad \text{for } q = 0, 1, \ldots, h$$

where N_x is odd if det $R(i_1i_2\cdots i_q; j_1j_2\cdots j_q)$ and det $H(i_1i_2\cdots i_q; j_1j_2\cdots j_q)$ are of opposite sign. Otherwise N_x is even. $h = \text{Min}\left\{\text{rank of either } \overline{R_p(k_1k_2\cdots k_{v-1})} \text{ or } \overline{H_p(k_1k_2\cdots k_{v-1})}\right\}$

Corollary 9: Let $F = R \cap H$. Then

$$T_1 = \left\{ (-1)^{N_1 + N_x} \quad k_1k_2\cdots k_{v-1} \frac{j}{i} \right\}$$

where $F(i; j)$ is a non-singular submatrix of order 1 (non-zero entry) in $\overline{F_p(k_1k_2\cdots k_{v-1})}$, and N_x is even or odd according to whether the entry in $F(i; j)$ is 1 or -1.

It should be noted that, for the enumeration of the corresponding non-singular submatrices of order $\geqslant 2$ in $\overline{R_p(k_1k_2\cdots k_{v-1})}$ and $\overline{H_p(k_1k_2\cdots k_{v-1})}$, the whole process outlined in Corollary 9 may be repeated.

Similarly, let $A^{i*}_{n_tn_r}$ and $A^{v*}_{n_un_r}$ be used to denote the matrices $A^{i}_{n_tn_r}$ and $A^{v}_{n_un_r}$ in proper form, respectively, such that their defining major submatrices consist of the same columns. Assume W is the matrix obtained from $A^{v*}_{n_un_r}$ by row interchanges and multiplying rows by (-1) if necessary such that W and $A^{i*}_{n_tn_r}$ have the same defining major submatrices. The number of such row interchanges plus the number of (-1)' s needed is represented by N_2. For convenience, the symbol $t^{n_t,n_r}_{n_u,n_r}$ is used to denote a complete two-tree of the graph such that it is a two-tree separating nodes n_t and n_r in the current graph, and again it is a two-tree separating nodes n_u and n_r in the voltage graph.

Theorem 8: With the same notation used above, the corresponding non-singular submatrices $\overline{Z_p(k_1k_2\cdots k_{v-2})}$ and $\overline{W_p(k_1k_2\cdots k_{v-2})}$ are in one-to-one correspondence with the complete two-trees of the form $t^{n_t,n_r}_{n_u,n_r}$ of G where $Z = A^{i*}_{n_tn_r}$. Furthermore, the set $T^1_{n_tn_u,n_r}$ of all distinct complete two-trees of distance 1 from the reference two-tree $t^{o}_{n_tn_u,n_r} = k_1k_2\cdots k_{v-2}$ is given by

$$T^1_{n_tn_u,n_r} = \left\{ (-1)^{N_2 + N_y} \quad k_1k_2\cdots k_{v-2} \frac{j}{i} \right\}$$

where $X(i; j)$ is a non-singular submatrix of order 1 in $\overline{X_p(k_1k_2\cdots k_{v-2})}$, and $X = Z \cap W$, and N_y is even or odd according to whether the entry in $X(i; j)$ is 1 or -1.

Note that F and X are matrices already in proper form. If the reference tree and two-tree consist only of ordinary elements, then the sign is known to be positive. This means that we may greatly simplify the procedure by picking the defining major submatrix corresponding to a complete tree (or two-tree) consisting only of ordinary elements of the network.

The results discussed above for the incidence matrices of a graph can easily be extended to the cutset matrices of a graph.

It should also be noted that if Theorem 8 is used to evaluate the cofactor of the (k,j)-element of the node-admittance determinant, then each term of the set of complete two-trees obtained by Theorem 8 must be multiplied by $(-1)^{k+j}$. Otherwise, the set of complete two-tree products only represents the minor of the (k,j)-element of the node-admittance determinant.

6. THE ENUMERATION OF DIRECTED TREES AND DIRECTED TWO-TREES OF A GRAPH WITHOUT DUPLICATIONS

Recently, a method of evaluating general network determinants and their cofactors by means of directed trees and directed two-trees is proposed by Chen[8] which shows considerable improvement over the older ones.[2,3,9] For a more extensive treatment, reference should be made to the various papers and works in the field.[8,10] In the following we shall present a method of enumerating all directed trees and directed two-trees of a graph without duplications.

Let G be a labelled directed linear graph of v nodes and e edges, and A^+ be the matrix obtained from A by replacing entries -1 of A by 0 where A again is the complete incidence matrix of G. Also let R and H be used to denote the matrices A_{n_r} and $A^+_{n_r}$ in proper form, respectively, such that their defining major submatrices are identical (by interchanging rows and multiplying rows by (-1) if necessary). As usual, the subscripts of the matrices are used to indicate that the rows corresponding to the subscripts have been deleted.

Theorem 9: With the same notation used above, the corresponding non-singular submatrices of order q in both $\overline{H_p(k_1k_2\ldots k_{v-1})}$ and $\overline{R_p(k_1k_2\ldots k_{v-1})}$ are in one-to-one correspondence with the directed trees t_{n_r} of distance q from the reference directed tree $t^o_{n_r}$ where t_{n_r} is the symbol for a directed tree with reference node n_r, and $q = 0, 1, \ldots, h$ where $h = \text{Min}\left[\text{rank of either } \overline{H_p(k_1k_2\ldots k_{v-1})} \text{ or } \overline{R_p(k_1k_2\ldots k_{v-1})}\right]$, and $t^o_{n_r} = k_1k_2\ldots k_{v-1}$.

Proof: Since directed trees of G are in one-to-one correspondence with the non-zero products of the corresponding major determinants of $A^+_{n_r}$ and the transpose of A_{n_r}, the theorem follows immediately.

Corollary 10: Let $T^1_{n_r}$ be the set of directed trees of distance 1 from $t^o_{n_r}$ in G. Then

$$T^1_{n_r} = \left\{ k_1k_2\ldots k_{v-1} \frac{i}{i} \right\}$$

where F(i;j) is a non-singular submatrix of order 1 in $\overline{F_p(k_1k_2\ldots k_{v-1})}$, and $F = R \cap H$.

Similarly, let W and Z be used to denote the matrices $A^+_{n_in_r}$ and $A_{n_jn_r}$ in proper form, respectively, such that their defining major submatrices are identical. For convenience, $t_{n_in_j,n_r}$ is the symbol for a directed two-tree in which the nodes n_i and n_j are in one component; the node n_r in another component, and nodes n_i and n_r (the first subscripts) are the reference nodes of the components.

Theorem 10: With the same notation used above, the corresponding non-singular submatrices of order q in both $\overline{W_p(k_1k_2\ldots k_{v-2})}$ and $\overline{Z_p(k_1k_2\ldots k_{v-2})}$ are in one-to-one correspondence with the directed two-trees of the form $t_{n_in_j,n_r}$ of distance q from the reference directed two-tree $t^o_{n_in_j,n_r} = k_1k_2\ldots k_{v-2}$, for $q = 0, 1, \ldots, w$ where $w = \text{Min}\left\{\text{rank of either } \overline{W_p(k_1k_2\ldots k_{v-2})} \text{ or } \overline{Z_p(k_1k_2\ldots k_{v-2})}\right\}$. Furthermore, the set $T^1_{n_in_j,n_r}$ of directed two-trees $t_{n_in_j,n_r}$ of distance 1 from $t^o_{n_in_j,n_r}$ is given by

$$T^1_{n_in_j,n_r} = \left\{ k_1k_2\ldots k_{v-2} \frac{i}{i} \right\}$$

where X(i; j) is a non-singular submatrix of order 1 in $\overline{X_p(k_1k_2\ldots k_{v-2})}$, and $X = W \cap Z$.

It should be noted that, by the repeated application of the technique of generating directed trees (two-trees) of distance 1 from the reference directed tree (two-tree), the set of directed trees (two-trees) of distance $\geq$ 2 can be obtained.

7. CONCLUSIONS

Methods for listing all possible trees, two-trees, complete trees, complete two-trees, directed trees and directed two-trees without duplications are presented. These methods are readily adaptable for digital computation, and are easy to remember and simple to operate.

Unlike other techniques, the procedure presented here can also be used to generate the dual subgraphs. For example, the method can be used to generate all possible cotrees and sign of cotree determinants for mesh-basis analysis. On the other hand, the dual procedure of other techniques, such as the one recently given by Mayeda and Seshu,[1] is by no means trivial.

REFERENCES

1. W. Mayeda, and S. Seshu, "Generation of Trees Without Duplications," IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 181-185, June, 1965.

2. C. L. Coates, "General Topological Formulas for Linear Network Functions" IRE Trans. on Circuit Theory, Vol. CT-5, pp. 42-54, March, 1958.

3. W. Mayeda, "Topological Formulas for Active Networks," Int. Tech. Rept., No. 8, U.S. Army Contract No. DA-11-022-ORD-1983, Univ. of Illinois, Urbana, Ill., January, 1958.

4. T. Fujisawa, "On a Problem of Network Topology," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 261-266, September, 1959.

5. H. Watanabe, "A Computational Method for Network Topology," IRE Trans. on Circuit Theory, Vol. CT-7, pp. 296-302, September, 1960.

6. S. L. Hakimi, "On Trees of a Graph and Their Generation," Jour. Franklin Institute, Vol. 272, pp. 347-359, Nov., 1961.

7. S. Seshu, and M. B. Reed, "Linear Graphs and Electrical Networks," Addison-Wesley Publishing Company, Inc., Reading, Mass., 1961.

8. W. K. Chen, "Topological Analysis for Active Networks," IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 85-91, March, 1965.

9. S. J. Mason, "Topological Analysis of Linear Non-reciprocal Networks," Proc. IRE, Vol. 45, pp. 829-838, June 1957.

10. A. Talbot, "Topological Analysis for Acive Networks " IEEE Trans. on Circuit Theory, Vol. CT-13, pp. 111-112, March, 1966.

11. H. M. Trent, "A Note on the Enumeration and Listing of All Possible Trees in a Connected Linear Graph, " Proc. Natl. Acad. Sciences, Vol. 40, pp. 1004-1007, October, 1954.

12. S.L. Hakimi, and D. G. Green, "Generation and Realization of Trees and k-Trees." IEEE Trans. on Circuit Theory, Vol. CT-11, pp. 247-255, June, 1964.

13. W. Mayeda, "Reducing Computation Time in the Analysis of Networks by Digital Computer," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 136-137, March, 1959.

14. H. Watanabe, "A Method of Tree Expansion in Network Topology," IRE Trans. on Circuit Theory, Vol. CT-8, pp. 4-10, March, 1961.

15. N. Nakagawa, "On Evaluation of the Graph Trees and the Driving Point Admittance," IRE Trans. on Circuit Theory, Vol. CT-5, pp. 122-127, June, 1958.

16. W. K. Chen, "Generation of Trees and k-Trees," Proc. Third Allerton Conference on Circuit and System Theory, pp. 889-899, October, 1965.

A GRAPH-THEORY APPROACH TO THE ANALYSIS OF MULTITERMINAL ELEMENT NETWORKS

MIRKO M. MILIĆ
Electrical Engineering Department, Imperial College, London, England.[*]

ABSTRACT

A graph-theory method is developed for the analysis of multiterminal-element networks. As contrasted with other existing graph-theory methods which are all based on the node-equilibrium equations, the proposed method introduces a new graph for representing the mesh-equilibrium equations of planar networks. When the network configuration is such that the mesh analysis is better suited than the node analysis, this method involves a lesser amount of work in comparison with the existing methods. In addition, mutually coupled inductances are easier to handle and no assumption of non-tight coupling is required.

I. INTRODUCTION

One of the main problems in the topological analysis of general linear networks is the finding of a suitable graph representation for multiterminal elements (m.t.e.). Percival (1) introduced the concept of a mathematical equivalent circuit for representing an m.t.e. described by the indefinite admittance matrix. Coates (2) and Mayeda (3) stated topological formulae for the analysis of general networks on the node basis. Lately, several topological procedures for calculating the determinant and cofactors of the node admittance matrix have been developed using a simpler graph (4,5). However, no topological method on the mesh basis seems to have been developed. The present paper deals with such a method. In the second part, a procedure is described for associating a graph with an indefinite immittance matrix. The third part deals with topological formulae on the mesh basis for the short-circuit functions of a 2-port.

II. GRAPHICAL REPRESENTATION OF MULTITERMINAL ELEMENTS[**]

Consider the N-terminal element M without internal sources in Fig. 1. It may be described by its indefinite immittance matrix $\underline{w}_{(N)}$ which relates the unknown variables vector $\underline{x}_{(N)}$ (terminal voltages v_k or mesh currents i_k) to the forcing variables vector $\underline{f}_{(N)}$ (terminal currents j_k or mesh voltages e_k) by the equation

$$\underline{w}_{(N)}\,\underline{x}_{(N)} = \underline{f}_{(N)} \quad . \tag{1}$$

An m.t.e. network consists of a finite set of interconnected m.t.elements and sources (Fig. 2). The terminals common to two or more m.t.e.'s are the nodes ν_k $(k = 1,2,\ldots,c)$. If the network is planar, the regions common to two or more m.t.e.'s are the meshes μ_k $(k = 1,2,\ldots,p)$. We assume the network to be planar and connected and assign a clockwise orientation to all finite region meshes and a counterclockwise orientation to the infinite region mesh (the outside mesh). The complete set of node or mesh equations is

$$\underline{W}\,\underline{x} = \underline{f} \tag{2}$$

---

[*] On leave of absence from the Faculty of Electrical Engineering, University of Belgrade, Belgrade, Yugoslavia.

[**] To meet the limitations of length imposed on authors, the details of analysis have had to be omitted, but a fuller analysis will be published later.

where $\underline{x}$ is the column matrix of the network variables (node voltages $\underline{v}$ or mesh currents $\underline{i}$), $\underline{f}$ is the column matrix of sources (node current sources $\underline{j}$ or mesh voltage sources $\underline{e}$) and $\underline{W}$ is the network indefinite immittance matrix (indefinite admittance matrix $\underline{Y}$ or indefinite impedance matrix $\underline{Z}$). Matrix $\underline{W}$ is related to the component m.t.e. indefinite immittance matrices $\underline{W}^{(k)}$ as

$$\underline{W} = \sum_k \underline{W}^{(k)} \qquad (3)$$

where $\underline{W}^{(k)}$ is derived from $\underline{w}_{(k)}$ by relabelling its rows and columns to correspond to the indices of the network variables shared by the m.t.e. M_k (6). The relabelled rows and columns are now the rows and columns of a c x c (p x p) matrix, the remaining elements of which are zeros.

In order to associate a topological structure with the complete set of equilibrium equations (2), we decompose each matrix $\underline{W}^{(k)}$ in eq. (3) in a unique way into a sum of component indefinite matrices such that each component matrix contains exactly one parameter. Thus one can imagine that eq. (2) has originated by summing up terms of the form

$$\begin{array}{cc} & \begin{array}{cccccccc} \quad & \quad & c & \quad & \quad & d & \quad & \quad \end{array} \\ \begin{array}{c} \\ \\ a \\ \\ b \\ \\ \\ \end{array} & \left[\begin{array}{cccccccc} & & 0 & & & 0 & & \\ & & \vdots & & & \vdots & & \\ \cdot & \cdot & w_k & \cdot & \cdot & -w_k & \cdot & \cdot \\ & & \vdots & & & \vdots & & \\ \cdot & \cdot & -w_k & \cdot & \cdot & w_k & \cdot & \cdot \\ & & \vdots & & & \vdots & & \\ & & 0 & & & 0 & & \end{array}\right] \end{array} \left[\begin{array}{c} 0 \\ \vdots \\ x_c \\ \vdots \\ x_d \\ \vdots \\ 0 \end{array}\right] = \left[\begin{array}{c} 0 \\ \vdots \\ f_a^{(k)} \\ \vdots \\ f_b^{(k)} \\ \vdots \\ 0 \end{array}\right] \qquad (4)$$

with $f_a^{(k)} = - f_b^{(k)}$. Now eq. (4) may be thought of as being the equation of the simplest m.t.e. referred to one of the systems of network variables. This m.t.e. may be looked upon as the basic network building block, and on the node basis it appears to be the same as Percival's mathematical network element. In the case a = c, b = d it stands for a two-terminal element.

On the node basis eq. (4) may be put into the form

$$j_{ab} = y_{ac,bd} v_{cd} \quad (j_{ab} = j_a = -j_b;\ v_{cd} = v_c - v_d) \qquad (5)$$

and represented by an oriented edge-pair E_k shown in Fig. 3. The black-arrow edge (the j-edge) denotes the current j_{ab} caused by the voltage v_{cd} which is represented by the white-arrow edge (the v-edge). The weight of both edges is the transadmittance of the edge-pair, $y_{ac,bd}$.

On the mesh basis eq. (4) reads

$$e_{ab} = z_{ac,bd} i_{cd} \quad (e_{ab} = e_a = -e_b;\ i_{cd} = i_c - i_d) \qquad (6)$$

and is represented by the oriented edge-pair E'_k shown in Fig. 4. The black-half-arrow edge (the e-edge) denotes the amount of voltage rise e_a in mesh μ_a (voltage drop e_b in mesh μ_b) produced by the current i_{cd} in the corresponding white half-arrow edge (the i-edge) which is the boundary between meshes μ_c and μ_d. The weight of both edges is the transimpedance of the edge-pair, $z_{ac,bd}$.

Denoting by $\underline{W}^{(k)}_{ac,bd}$ the square matrix in eq. (4), eq. (3) reads

$$\underline{W} = \sum_k \underline{W}^{(k)}_{ac,bd} \qquad (7)$$

The graphical counterpart of eq. (7) is the operation of forming the network graph from the component graphs E_k or E'_k. On the node basis, the network graph G_n is obtained by adding j-edges and v-edges between the network nodes the indices of which correspond to identical indices of rows and columns of the respective $Y^{(k)}_{ac,bd}$ matrices. The $\underline{G_n\text{-graph}}$ is then $G_n = G_j + G_v$, where the $\underline{G_j\text{-graph}}$ and the $\underline{G_v\text{-graph}}$ are the graphs constituted from the j-edges and v-edges, respectively. On the mesh basis, the network graph G_m is formed by inserting e-edges and i-edges into the network meshes the indices of which correspond to identical indices of rows and columns of the respective $Z^{(k)}_{ac,bd}$ matrices. The $\underline{G_m\text{-graph}}$ is then $G_m = G_e + G_i$, where the $\underline{G_e\text{-graph}}$ and $\underline{G_i\text{-graph}}$ are the graphs constituted from the e-edges and i-edges, respectively.

Notice that in forming the G_m-graph the total number of nodes c remains unchanged and that in forming the G_m-graph the total number of meshes p remains unchanged.

For a 2-terminal element edge-pair $j_{ab} = i_{ab}$ and $e_{ab} = v_{ab}$. It follows that $y_{aa,bb} \cdot z_{aa,bb} = 1$, and that the two graphs G_n and G_m of a 2-terminal element network are identical except for the edge labelling.

In Table 1 the G_n-graph and G_m-graph of some common m.t.e.'s are shown together with their indefinite matrices Y and Z. Figs. 6 and 7 show the G_n-graph and the G_m-graph respectively of the network of Fig. 5.

III. TOPOLOGICAL FORMULAE ON THE MESH BASIS.

Consider the 2-port in Fig. 8 in which each part is coupled with two meshes. The S.C. admittance matrix is

$$\underline{y} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \frac{1}{D} \begin{bmatrix} D_{(pq)(pq)} & D_{(rs)(pq)} \\ D_{(pq)(rs)} & D_{(rs)(rs)} \end{bmatrix} \qquad (8)$$

where D is the determinant of the mesh impedance matrix Z_o (Z_o is obtained from the indefinite impedance matrix Z of eq. (4) by deleting the row and column corresponding to the outside mesh of the network) and $D_{(pq)(rs)}$ is the compound cofactor of the (p,r)-element of D (6).

The mesh impedance matrix Z_o can be put in a form similar to that used with 2-terminal element networks. Let the G_m-graph have m finite meshes (m=p-1) and b edge-pairs. Then both the G_e-graph and the G_i-graph have each m finite meshes and b edges. The G_e-graph (G_i-graph) is obtained from the G_m-graph by contracting all the i-edges (e-edges). Now on each of the two graphs mesh incidence matrices are defined as follows. The mesh matrix B_e (B_i) of the G_e-graph (G_i-graph) is of order m x b with rows and columns numbered in the same order. For an edge-pair β_j of transimpedance $Z_{ac,bd}$ there is a 1 entry in column j and row a and a -1 entry in column j and row b of the matrix B_e, and a 1 entry in column j and row c and -1 entry in column j and row d in matrix B_i. For a 2-terminal element edge β_h of impedance $Z_{aa,bb}$ there is a 1 entry in column h and row a, and a -1 entry in column h and row b in both matrices.

If we denote by Z_b the b x b diagonal matrix of the transimpedances of the edge-pairs written in the same order as the columns of B_e and B_i, then it is easy to show that

$$Z_o = B_e Z_b B_i^t . \qquad (9)$$

Now we define a common cotree to be a set of edges which constitute a cotree both in the G_e-graph and in the G_i-graph. The common cotree product is the product of the transimpedances of a common cotree. The product of majors $|B_e^c|$ and $|B_i^c|$ corresponding to a common cotree is the sign of the common cotree.

Expansion of $D = \det Z_o$ by the Binet-Cauchy theorem gives

$$D = \sum_k \delta_k \lambda_k \qquad (10)$$

where λ_k is the k-th common cotree product and $\delta_k = |B_e^c||B_i^c|$ is the sign of the k-th common cotree. Since the structure of a cotree determinant is the same as the structure of a tree determinant, a dual procedure (3) can be employed to transform a given cotree determinant to a simpler equivalent form from which it is easier to reveal the changes necessary to make $|B_e^c|$ and $|B_i^c|$ coincide.

To this end, we notice that it is possible to associate with each mesh μ_p a unique chord β_k which defines a fundamental mesh μ_p^o encircling mesh μ_p. This chord will be called the principal chord of the mesh μ_p. The algorithm for determining δ_k is most easily formulated by defining the Cotree Permutation Schedule (C.P.S.). The C.P.S. consists of two rows and m columns. The first row refers to the chords of G_e and the second to the chords of G_i. Both sets of chords constitute a common cotree of G_m. In each row principal chords are listed in columns corresponding to the meshes to which they are pertinent and a minus superscript is put on each principal chord the orientation of which does not coincide with the orientation of its pertinent mesh.

Let s be the total number of minus superscripts and t the total number of transpositions in one of the rows of the C.P.S. required to make the two rows identical. The sign of the k-th common cotree is then

$$\delta_k = (-1)^{s+t} . \qquad (11)$$

The application of this algorithm is greatly simplified by observing that δ_k can still be obtained if all the 2-terminal element chords of the k-th common cotree are opened and the C.P.S. is written only for the edge-pair chords of the remaining meshes. Particularly, if all chords are 2-terminal elements, $\delta_k = 1$, i.e. no C.P.S. is required.

As an example, consider the five-mesh G_m-graph shown in Fig. 9, where the G_e-graph and the G_i-graph are shown separately. Since β_1 and β_3 are 2-terminal element chords of a common cotree, they may be erased leaving a three-mesh graph whose C.P.S. is

$$\begin{array}{c|ccc} & \mu_3 & \mu_4 & \mu_5 \\ \hline G_e & \beta_5 & \beta_4 & \beta_2 \\ G_i & \beta_4 & \beta_2 & \beta_5 \end{array}$$

Here $s = 1$, $t = 2$. Hence $\delta_k = (-1)^{1+2} = -1$.

The evaluation of all the cofactors in eq. (8) can be reduced to the evaluation of the mesh determinants of the modified network. The modification of the network consists in representing the equation defining the network function by an edge-pair which is then incorporated in the network graph. For example, y_{12} can be found by inserting an e-edge of weight y_{12}^{-1} in the output mesh of the given network N (Fig. 10) and determining the signed common cotree products of the resulting network N_1. If the sum of the signed common cotree products is set to zero, a linear equation in y_{12} results from which y_{12} is obtained. Similarly, y_{11} is obtained by inserting a 2-terminal element edge of impedance y_{11}^{-1} in the input mesh as shown in Fig. 11 and by proceeding as above.

Example. The above technique will be described by the following example. Consider the network of Fig. 5 whose G_m-graph in Fig. 7 is redrawn in Fig. 12 for convenience. We find y_{21} and y_{11} with respect to the input mesh pair (1,4) and the output mesh pair (3,4). In order to find y_{21}, an e-edge is inserted between meshes 1 and 4 and an i-edge between meshes 3 and 4, both edges being of impedance $z = y_{21}^{-1}$. This edges are shown as dotted lines in Fig. 12a.

The sum of cotree products due to 2-terminal-element edges is found to be $D' = Z_d(Z_aZ_b + Z_aZ_c + Z_bZ_c)$.

The signed common cotree products containing at least one non-two-terminal element edge are: $-r_mZ_aZ_d$; $-Z'_{34}Z''_{34}Z_a$; $-Z'_{34}Z''_{34}Z_b$; $-zZ_bZ''_{34}$; $+zr_mZ''_{34}$. The signs of these products are determined by the C.P.S. The mesh determinant is $D = D' + D''$, where $D'' = -Z_aZ_dr_m-(Z_a+ Z_b)Z_{34}^2$.

$D_{(14)(14)}$ is easily found as the sum of the signed common cotree products containing Z_a with Z_a=1, i.e. $D_{(14)(14)} = (Z_b+ Z_c-r_m)Z_d-Z_{34}^2$. Finally, $D_{(14)(43)}$ equals the sum of the signed common cotree products containing z with $z = -1$. The required functions are $y_{21}=D_{(14(43)}/D$ and $y_{11}=D_{(14)(14)}/D$.

IV. CONCLUSIONS

It was shown that with every planar m.t.e. network which possesses both an indefinite admittance and an indefinite impedance matrix two different graphs may be associated, one of which represents the complete system of node equations and the other the complete system of mesh equations. It is only for two-terminal-element networks that these two graphs become identical, the edge weights of one of them being then the reciprocals of the corresponding edges of the other. If the network under consideration possesses only one type of indefinite matrix then only one graph exists. A typical situation of this sort is a network containing tightly coupled coils for which no Y matrix exists and hence no G_n-graph.

A procedure for associating an indefinite matrix with a graph was described and a dual algorithm for finding the sign of a common cotree product was given. For planar network configurations better suited for the mesh analysis than for the node analysis, the G_m-graph is simpler than the G_n-graph and topological formulae on the mesh basis involve less sign determinations than the formulae on the node basis do. (Only 3 sign determinations are needed for the mesh determinant of the network in Fig. 5, compared with 14 for the node determinant).

VI. REFERENCES.

(1) W. S. Percival, "Improved Matrix and Determinant Methods of Solving Networks". Proc. IEE (London), pt.IV, vol.101, pp.258 265.

(2) C. L. Coates, "General Topological Formulas for Linear Network Functions", IRE Trans. on Circuit Theory, vol.CT-5, pp.42-54, 1958.

(3) S. Seshu and M. B. Reed", "Linear Graphs and Electrical Networks", Addison-Wesley, Reading, Mass., 1961.

(4) A. Talbot, "Topological Analysis of General Linear Networks", Proc. 6th Midwest Symp. on Circuit Theory, Univ. of Wisconsin, May, 1963.

(5) A. Natham, "Topological Rules for Linear Networks", The Technion, Israel Inst. of Technology, Haifa, Publ.no. 21A, Oct.-Nov.,1963.

(6) V. P. Sigorski, "Methods of Analysis of Multiterminal-Element Networks", Kiev, Acad.Sc.,USSR, 1958 (in Russian).

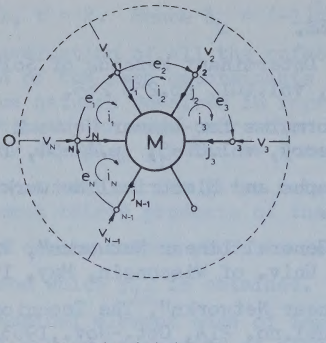

M
O
V_1
V_2
V_3
V_N
V_{N-1}
e_1
e_2
e_3
e_N
i_1
i_2
i_3
i_N
j_1
j_2
j_3
j_N
j_{N-1}
1
2
3
N
N-1

Fig.1 An N–terminal element

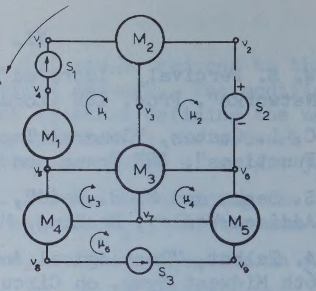

M_1
M_2
M_3
M_4
M_5
S_1
S_2
S_3
v_1
v_2
v_3
v_4
v_5
v_6
v_7
v_8
v_9
μ_1
μ_2
μ_3
μ_4
μ_5
μ_6

Fig.2 Network consisting of 5 m.t. elements, 2 current sources and 1 voltage source.

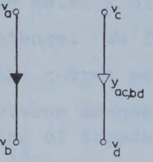

v_a
v_b
v_c
v_d
$y_{ac,bd}$
$y_{ac,bd}$

Fig.3 Graph of the equation

$j_{ab} = y_{ac,bd}\, v_{cd}$

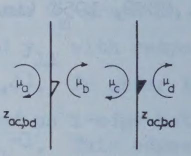

μ_a
μ_b
μ_c
μ_d
$z_{ac,bd}$
$z_{ac,bd}$

Fig.4 Graph of the equation

$e_{ab} = z_{ac,bd}\, i_{cd}$

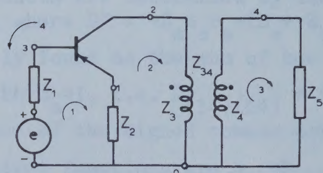

1
2
3
4
0
Z_1
Z_2
Z_3
Z_4
Z_5
Z_{34}
e
+
−

Fig.5

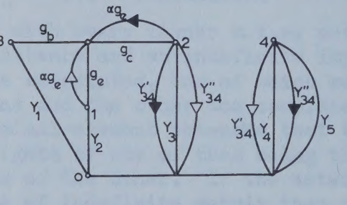

αg_c
g_b
g_c
g_e
αg_e
Y_1
Y_2
Y_3
Y_4
Y_5
Y'_{34}
Y''_{34}
Y'_{34}
Y''_{34}
0
1
2
3
4

Fig.6 G_n-graph of the network of Fig.5

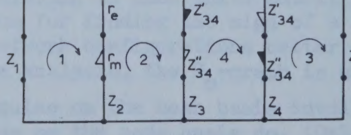

r_b
r_c
r_m
r_e
r_m
Z'_{34}
Z''_{34}
Z'_{34}
Z''_{34}
Z_1
Z_2
Z_3
Z_4
Z_5
1
2
3
4

Fig.7 G_m-graph of the network of Fig.5

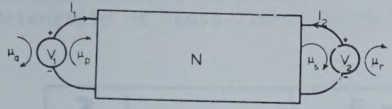

I_1
I_2
μ_a
V_1
μ_p
N
μ_s
V_2
μ_r
+
−
+
−

Fig.8 A general two port inscribed in a network

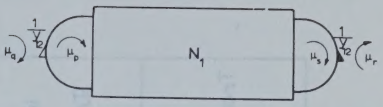

1/y_12
μ_a
μ_p
N_1
μ_s
1/y_12
μ_r

Fig.10 Network for evaluation of y_{12}

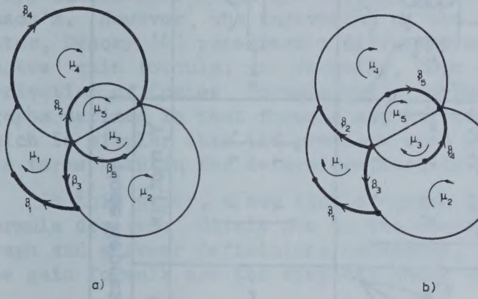

β_4
μ_4
β_2
μ_5
μ_3
μ_1
β_3
β_5
μ_2
β_1
μ_4
β_5
β_2
μ_5
μ_3
β_4
μ_1
β_3
μ_2
β_1
a)
b)

Fig.9 A five-mesh G_m-graph a) G_e-graph, b) G_i-graph

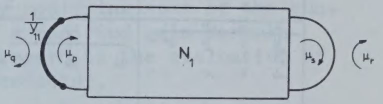
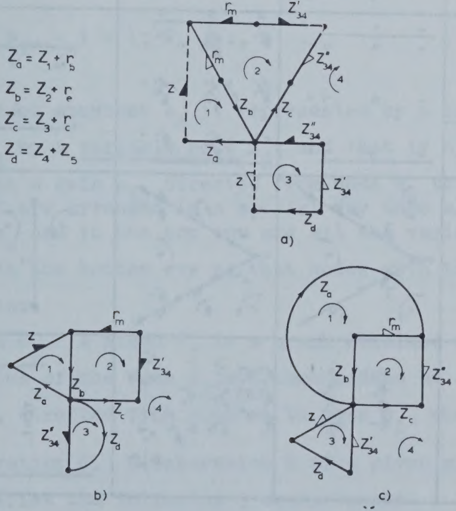

1/y_11
μ_a
μ_p
N_1
μ_s
μ_r

Fig 11 Network for evaluation of y_{11}

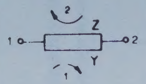
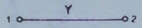
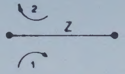
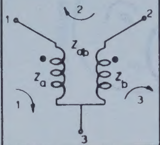
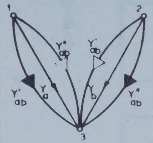
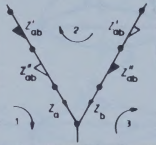
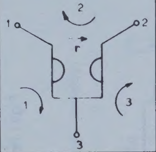
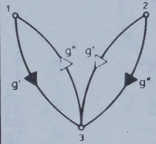
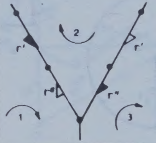
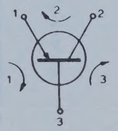
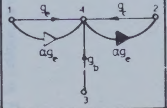
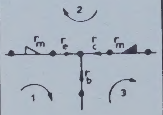


$$Z_a = Z_1 + r_b$$
$$Z_b = Z_2 + r_e$$
$$Z_c = Z_3 + r_c$$
$$Z_d = Z_4 + Z_5$$

r_m
Z'_34
r_m
Z''_34
2
4
z
1
Z_b
Z_c
Z''_34
Z_a
z
3
Z'_34
Z_d
a)
r_m
z
1
2
Z'_34
Z_b
Z_a
Z_c
4
Z'_34
3
Z_d
b)
Z_a
1
r_m
Z_b
2
Z''_34
z
Z_c
3
Z'_34
4
Z_d
c)

Fig.12 a) G_m-graph of the network of Fig.5 b) G_e-graph c) G_i-graph

TABLE 1

Multi-terminal element		Node Representation		Mesh Representation		Notes
		Y-matrix	G_n-graph	Z-matrix	G_m-graph	
Two-terminal element		$\begin{array}{c} & 1 & 2 \\ 1 & Y & -Y \\ 2 & -Y & Y \end{array}$		$\begin{array}{c} & 1 & 2 \\ 1 & Z & -Z \\ 2 & -Z & Z \end{array}$		$YZ=1$
Two mutually coupled coils	 1 2 2 Z_{ab} Z_a Z_b 1 3 3	$\begin{array}{c} & 1 & 2 & 3 \\ 1 & Y_1 & Y_{ab} & -Y_a-Y_{ab} \\ 2 & Y_{ab} & Y_b & -Y_b-Y_{ab} \\ 3 & -Y_a-Y_{ab} & -Y_b-Y_{ab} & Y_a+Y_b+2Y_{ab} \end{array}$	 1 2 Y''_{ab} Y'_{ab} Y'_{ab} Y_a Y_b Y''_{ab} 3	$\begin{array}{c} & 1 & 2 & 3 \\ 1 & Z_a & -Z_a+Z_{ab} & -Z_{ab} \\ 2 & -Z_a+Z_{ab} & Z_a+Z_b+2Z_{ab} & -Z_b+Z_{ab} \\ 3 & -Z_{ab} & -Z_b+Z_{ab} & Z_b \end{array}$	 Z'_{ab} 2 Z'_{ab} Z''_{ab} Z''_{ab} 1 Z_a Z_b 3	$Y_a=\frac{Z_b}{\Delta}$; $Y_b=\frac{Z_a}{\Delta}$ $Y_{ab}=\frac{-Z_{ab}}{\Delta}$ $\Delta=Z_aZ_b-Z_{ab}^2$ $Y'_{ab}=Y''_{ab}=Y_{ab}$ $Z'_{ab}=Z''_{ab}-Z_{ab}$
Ideal gyrator		$\begin{array}{c} & 1 & 2 & 3 \\ 1 & 0 & g & -g \\ 2 & -g & 0 & g \\ 3 & g & -g & 0 \end{array}$		$\begin{array}{c} & 1 & 2 & 3 \\ 1 & 0 & r & -r \\ 2 & -r & 0 & r \\ 3 & r & -r & 0 \end{array}$		$r'=r''=r=\frac{1}{g}$ $g'=g''=g=\frac{1}{r}$
Transistor		$\begin{array}{c} & 1 & 2 & 3 & 4 \\ 1 & g_e & 0 & 0 & -g_e \\ 2 & \alpha g_e & g_c & 0 & \alpha g_e-g_c \\ 3 & 0 & 0 & g_b & -g_b \\ 4 & \alpha g_e-g_e & -g_c & -g_b & g_e+g_c+g_b-\alpha g_e \end{array}$	 1 g_e 4 g_c 2 αg_e g_b αg_e 3	$\begin{array}{c} & 1 & 2 & 3 \\ 1 & r_e+r_b & -r_e & -r_b \\ 2 & r_m-r_e & r_e+r_c-r_m & -r_c \\ 3 & -r_b-r_m & r_m-r_c & r_b+r_c \end{array}$	 2 r_m r_e r_c r_m r_b 1 3	$g_er_e=1$; $g_br_b=1$ $g_cr_c=1$; $\alpha=g_cr_m$

A NEW TOPOLOGICAL TECHNIQUE FOR SOLVING A LINEAR SYSTEM

S. PARK CHAN AND HSIA-NUNG MAI
University of Santa Clara, Santa Clara, California

ABSTRACT

Although Mason's signal flow graph [1], [2] gives a very good insight of the casual relationships between the variables of the linear system under study, the derivation and the expansion of his general gain formula are quite complicated. Subsequently, Coates [3] defined another kind of graph called flow graph and derived a gain formula which is simpler than Mason's. However, the derivation of the gain formula is still not simple. Later, Desoer [4] presented a different and much simpler derivation of Caotes' gain formula; and recently, Chan and Bapna [5] simplified Desoer's derivation of Coates' formula still further by modifying some of the subgraphs defined in that formula and arrived at a modified gain formula which is simpler than the previous ones with regard to the evaluation of the terms both in the determinant and in the cofactor.

In this paper, a new kind of graph is defined and a different gain formula derived. Mainly due to the characteristics of the newly defined graph and a fewer definitions necessary, the derivation and expansion of the gain formula are the simplest among the existing ones.

DEFINITIONS

Definition 1: Flow graph G. A flowgraph G is an oriented graph representing a system of simultaneous linear algebraic equations

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 1, 2, \ldots, n \ .$$

in such a way that each constant b_i is represented by a constant node b_i and each variable x_j by a variable node x_j, and that if $a_{ij} \neq 0$, there is an oriented edge with a gain a_{ij} directed from node x_j to node b_i. Furthermore, the nodes of G are arranged in a special way that all the constant nodes $b_1, b_2, \ldots, b_n$ are in the top row and all the variable nodes $x_1, x_2, \ldots x_n$ are in the bottom row so that nodes with the same subscript are on a vertical line.

Definition 2: Graph G_j. A graph G_j is a graph obtained from G by deleting all the outgoing edges of the node x_j and then adding an oriented edge with a gain b_i if $b_i \neq 0$, directed from node x_j to node b_i, where $i = 1, 2, \ldots, n$.

Definition 3: Separation S. A separation S of a given graph (G or G_j) is a subgraph which satisfies the following 3 requirements:

(1) It contains all the nodes of the graph (G or G_j) ;

(2) Each variable node x_j has exactly one outgoing edge;

(3) Each constant node b_i has exactly one incoming edge.

Definition 4: Separation Product S_p. A separation product S_p is the product of the gains of all the edges in S. A separation product of G is denoted by $S_p(G)$ and that of G_j by $S_p(G_j)$.

LEMMAS AND THEOREMS

Consider a linear system of n independent algebraic equations in n unknowns

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 1, 2, \ldots, n .$$

which, in matrix notation, may be written as

$$[a_{ij}]\,[x] = [b]$$

where $[a_{ij}]$ is the coefficient matrix, $[x]$ is the unknown column matrix, and $[b]$ is the constant column matrix.

The solution of the variable x_j, using Cramer's rule, is given by

$$x_j = \frac{\Delta_j}{\Delta} \quad (\text{for } \Delta \neq 0)$$

where Δ is the determinant of the coefficient matrix $[a_{ij}]$, and Δ_j is the determinant obtained from Δ by replacing the j^{th} column with the constant column. That is,

$$\Delta = \begin{vmatrix} a_{11}a_{12} & \cdots & a_{1n} \\ a_{21}a_{22} & \cdots & a_{2n} \\ & \cdots & \\ a_{n1}a_{n2} & \cdots & a_{nn} \end{vmatrix} , \quad \Delta_j = \begin{vmatrix} a_{11}a_{12} & \cdots & a_{1,j-1}b_1a_{1,j+1} & \cdots & a_{1n} \\ a_{21}a_{22} & \cdots & a_{2,j-1}b_2a_{2,j+1} & \cdots & a_{2n} \\ & & \cdots & & \\ a_{n1}a_{n2} & \cdots & a_{n,j-1}b_na_{n,j+1} & \cdots & a_{nn} \end{vmatrix}$$

Lemma 1: Each term in the expansion of Δ corresponds to a separation product S_{p_j} of G.

Proof: Note that any product term T_j of Δ when expanded by rows, has the form

$$T_j = a_{ij_1}a_{2j_2} \cdots a_{nj_n} \quad (a_{ij_r}, \ i = 1, 2, \ldots, n \ ; \ j_1, j_2, \ldots, j_n \text{ is a permutation of the set } r = 1,2,\ldots,n)$$

where the first subscript i of a_{ij} identifies the row and the second subscript j indicates the column in which the element a_{ij} is located. As the expansion of the determinant is carried out in the ascending order of rows, the first subscript, i, of the elements of the product-term T_j is fixed; while each of the second subscripts, $j_1, j_2, \ldots, j_n$, takes a value from 1 to n, where no two values of j_r are equal. Therefore, T_j represents a possible permutation of the second subscripts, j_r, r = 1,2, ..., n, of a_{ij}.

The total number of terms in Δ is equal to the number of all possible permutations of the columns of $[a_{ij}]$. Now, in T_j there are n elements, each of which is in a distinct row and in a distinct column. The elements in row i of $[a_{ij}]$ represent the incoming edges at node b_i while the elements in column j correspond to the outgoing edges at node X_j of the flow graph G. Hence T_j corresponds to a subgraph of G which every constant node b_i has exactly one incoming edge and every variable node X_j has exactly one outgoing edge. Furthermore, T_j contains elements from each row and each column of $[a_{ij}]$. This means that the corresponding graph of T_j includes all the constant and variable nodes of G and hence satisfies all three requirements in the definition of a separation. Therefore, T_j corresponds to a separation product S_{p_j} of the flow graph G.

Lemma 2: Each separation product S_{p_j} of G corresponds to a term in the expansion of Δ .

Proof: In a flow graph G which corresponds to a linear system of n independent algebraic equations in n unknowns, there are n constant nodes and n variable nodes. In light of the definition of a separation, it is obvious the a separation consists of n edges and a separation product has n elements (n edge-gains). Consider any separation product S_{p_j} :

$$S_{p_k} = a_{1k_1} a_{2k_2} \dots a_{nk_n} \quad (a_{ik_r},\ i = 1, 2, \dots, n;\ k_1, k_2, \dots, k_n \text{ is a permutation of the set } r = 1, 2, \dots, n.)$$

where the first subscript i of a_{ik} indicates that the edge with a gain a_{ik} terminates at node b_i and the second subscript k identifies that the same edge originates from node X_k. If the n edges of the corresponding separation S_k are arranged so that they terminate at the n constant nodes in the order $b_1, b_2, \dots, b_n$, then the first subscript, i, of the elements of the separation product S_{p_k} is fixed while each of the second subscripts, $k_1, k_2, \dots, k_n$, takes a value from 1 to n, where no two values of k_r are equal. This is because in a separation, there is exactly one edge originating from each variable node X_j. Therefore S_{p_k} represents a possible permutation of the second subscripts, k_r, $r = 1, 2, \dots, n$, of a_{ik} which satisfies the definition of a term in the expansion of Δ ; and consequently each S_{p_k} of G corresponds to a term in the expansion of the determinant.

Combining Lemma 1 and Lemma 2, we have the conclusion as stated in Lemma 3.

Lemma 3: The set of all the product-terms in the expansion of Δ of $[a_{ij}]$ of a linear system of n independent algebraic equations is equal to the set of the separation-products of the corresponding flow graph G.

Lemma 4: The sign of the separation product S_{p_j} of a separation S_j of G is given by $(-1)^{D-1}$ for $D \neq 0$ and is equal to +1 for $D = 0$ where D is the number of non-vertical edges of S_j.

Proof: Recall a property of a determinant: If a product-term has all of its elements on the principal diagonal of the determinant, then the sign of that product-term is positive.

In a separation S_j, all the vertical edges correspond to the diagonal elements of Δ (= det $[a_{ij}]$) but all the non-vertical edges correspond the off-diagonal elements in the determinant as shown in Fig. 1. Thus, if all the edges of S_j are vertical, the sign of S_{p_j} is positive.

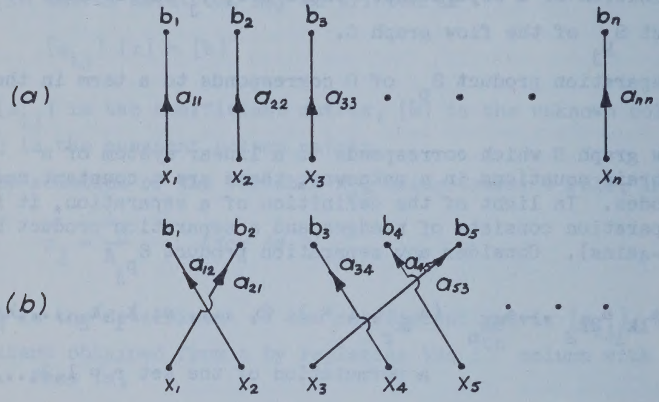


Fig. 1

(a) Vertical edges of a separation

(b) Non-vertical edges of a separation

Observe that each non-zero term in the expansion of Δ has exactly one element in each row and one element in each column. Correspondingly, each separation has exactly one edge between one of the constant node b_i and one of the variable node X_j.

Note also that for a non-zero term with two off-diagonal elements a_{ij} and a_{ji} $(i \neq j)$ in the expansion of Δ, it requires one interchange of columns (or rows) to place both elements on the principal diagonal of Δ as shown in Fig. 2.

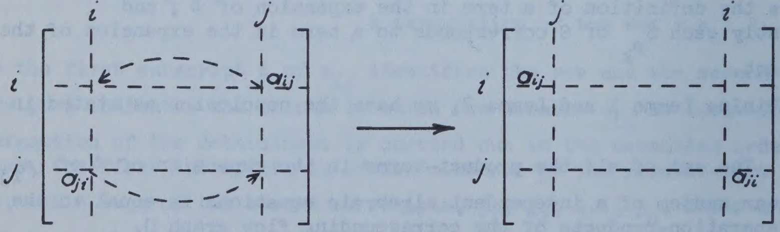
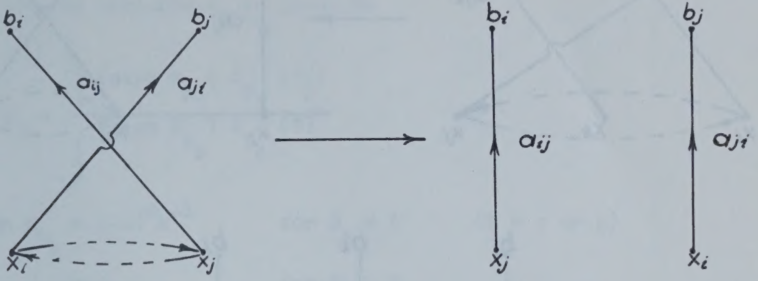


Fig. 2. One interchange of columns

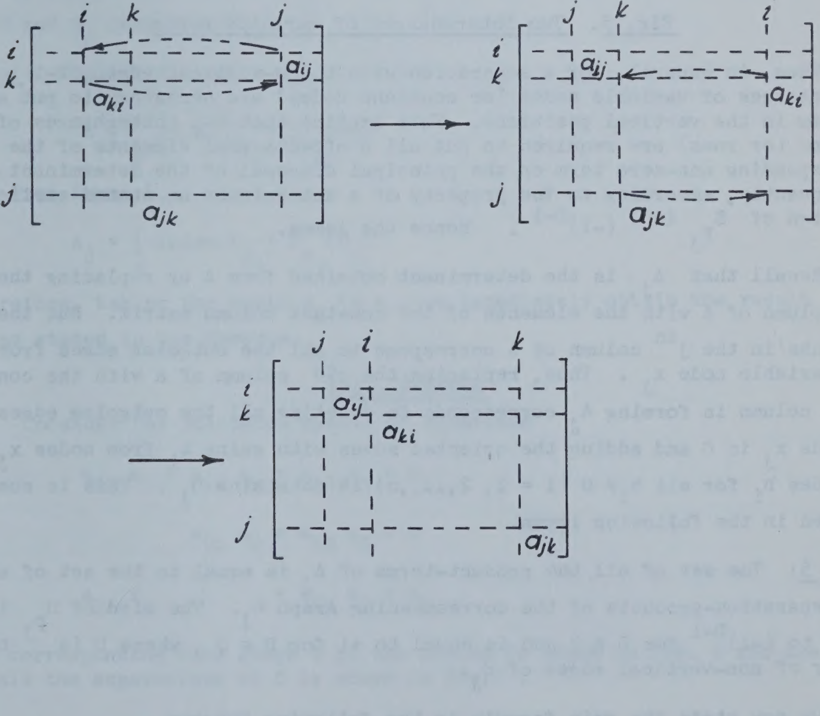
The equivalent operation for the corresponding separation in the graph G is the interchange of the variable nodes (or constant nodes) of the two non-vertical edges a_{ij} and a_{ji} $(i \neq j)$ of the separation so as to put them in the vertical position as illustrated in Fig. 3.



b_i
b_j
a_{ij}
a_{ji}
x_i
x_j
b_i
b_j
a_{ij}
a_{ji}
x_j
x_i

Fig. 3. One interchange of variable nodes

For a non-zero term with three off-diagonal elements in the expansion of Δ, two interchanges of columns (or rows) are required in order to put the off-diagonal elements on the principal diagonal (Fig. 4). Equivalently, it requires two interchanges of the variable nodes (or constant nodes) to put the three non-vertical edges of the corresponding separation in the vertical positions (Fig. 5).



i
k
j
i
k
j
a_{ij}
a_{ki}
a_{jk}
j
k
i
i
k
j
a_{ij}
a_{ki}
a_{jk}
j
i
k
i
k
j
a_{ij}
a_{ki}
a_{jk}

Fig. 4. Two interchanges of columns

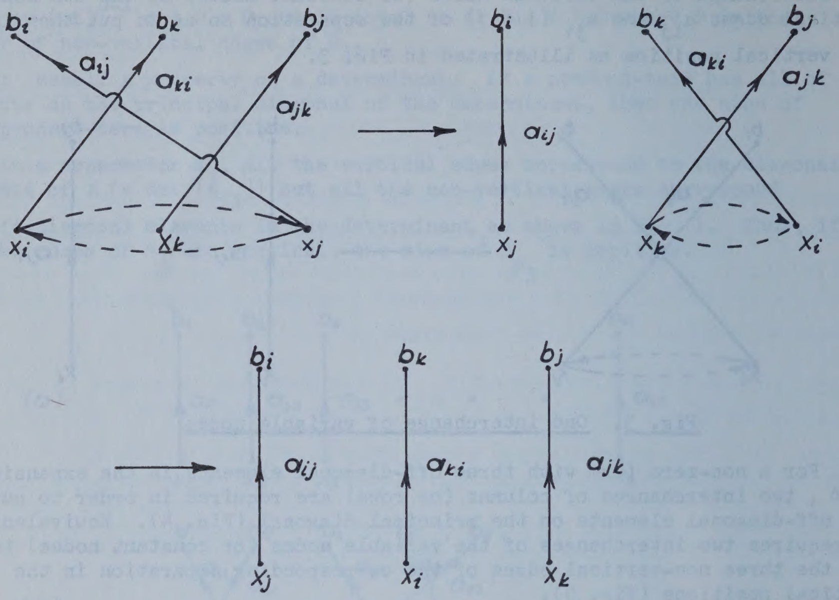


Fig. 5. Two interchanges of variable nodes

Thus, in general, for a separation with D non-vertical edges, D-1 interchanges of variable nodes (or constant nodes) are necessary to put all of them in the vertical positions. This implies that D-1 interchanges of columns (or rows) are required to put all D off-diagonal elements of the corresponding non-zero term on the principal diagonal of the determinant Δ. Consequently, according to the property of a determinant as stated earlier, the sign of S_{p_j} is $(-1)^{D-1}$. Hence the lemma.

Recall that Δ_j is the determinant obtained from Δ by replacing the j^{th} column of Δ with the elements of the constant column matrix. But the elements in the j^{th} column of Δ correspond to all the outgoing edges from the variable node x_j. Thus, replacing the j^{th} column of Δ with the constant column in forming Δ_j corresponds to deleting all the outgoing edges of node x_j in G and adding the oriented edges with gains b_i from nodes x_j to nodes b_i for all $b_i \neq 0$ $(i = 1, 2, \ldots, n)$ in obtaining G_j. This is summarized in the following lemma.

Lemma 5: The set of all the product-terms of Δ_j is equal to the set of all the separation-products of the corresponding graph G_j. The sign of S_{p_j} is equal to $(-1)^{D-1}$ for $D \neq 0$ and is equal to +1 for $D = 0$, where D is the number of non-vertical edges of S_j.

We now state the gain formula in the following theorem.

Theorem: In a linear system of n simultaneous independent algebraic equations in n unknowns

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \quad , \qquad i = 1,2,\ldots, n \ ,$$

the value of the variable x_j is given by

$$x_j = \frac{\sum_{\tau} (\text{sign } S_{p_\tau}) S_{p_\tau}(G_j)}{\sum_{\rho} (\text{sign } S_{p_\rho}) S_{p_\rho}(G)}$$

where $\text{sign } S_{p_k} = (-1)^{D_k - 1}$ for $D_k \neq 0$ $(k = \tau \text{ or } \rho)$

$= 1$ for $D_k = 0$

D_τ = the number of non-vertical edges in the τ^{th} separation of the flow graph G_j , $S_{p_\tau}(G_j)$= the separation-product of the τ^{th} separation of G_j , D_ρ = the number of non-vertical edges in the ρ^{th} separation of the graph G . $S_{p_\rho}(G)$ = the separation-product of the ρ^{th} separation of G . The summations of the separation-products $S_p(G)$ and $S_p(G_j)$ are taken over all separations of G and G_j respectively.

Proof: Combining Lemma 3 and 4, we get

$$\Delta = \sum_{\rho} (\text{sign } S_{p_\rho}) S_{p_\rho}(G)$$

and from Lemma 5,

$$\Delta_j = \sum_{\tau} (\text{sign } S_{p_\tau}) S_{p_\tau}(G_j)$$

Therefore, taking the ratio Δ_j to Δ , we immediately obtain the result for x_j as stated in the theorem.

III. EXAMPLE

Consider the following system of equations

$$\begin{aligned} a_{11} x_1 + a_{12} x_2 + a_{13} x_3 &= b_1 \\ a_{22} x_2 + a_{23} x_3 &= 0 \\ a_{31} x_1 + a_{33} x_3 &= b_3 \end{aligned}$$

The corresponding flow graph G of the system is shown in Fig. 6 and the set of all the separations of G is shown in Fig.7 .

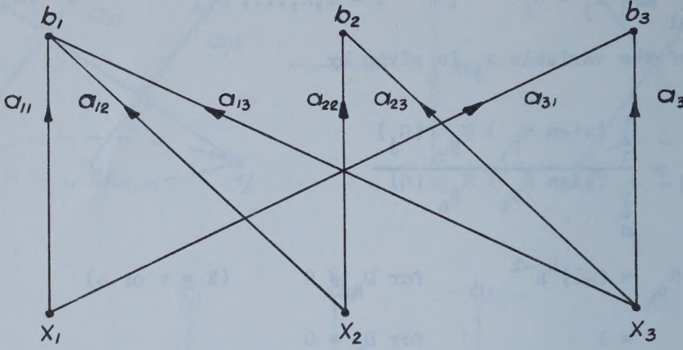

b_1
b_2
b_3
a_{11}
a_{12}
a_{13}
a_{22}
a_{23}
a_{31}
a_{33}
x_1
x_2
x_3

Fig. 6. The flow graph G of Ex. 1

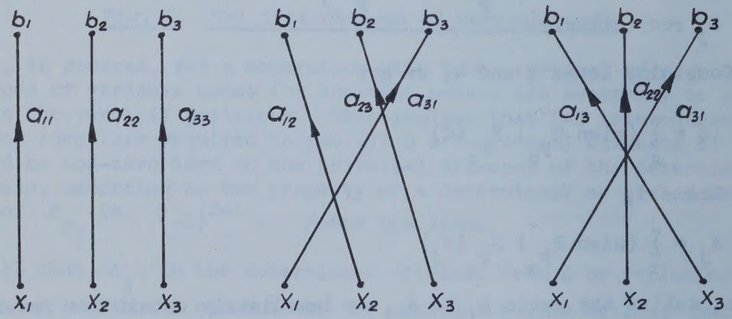

b_1
b_2
b_3
a_{11}
a_{22}
a_{33}
x_1
x_2
x_3
b_1
b_2
b_3
a_{12}
a_{23}
a_{31}
x_1
x_2
x_3
b_1
b_2
b_3
a_{13}
a_{22}
a_{31}
x_1
x_2
x_3

Fig. 7. The set of all the separations of G

Suppose we are interested in the value of x_2. Then in G we delete all the outgoing edges from the variable node x_2 and add one edge with a gain b_1 from node x_2 to node b_1 and another edge with a gain b_3 from node x_2 to node b_3. The resulting graph G_2 and the set of all the separations of G_2 are shown in Figs. 8 and 9 respectively.

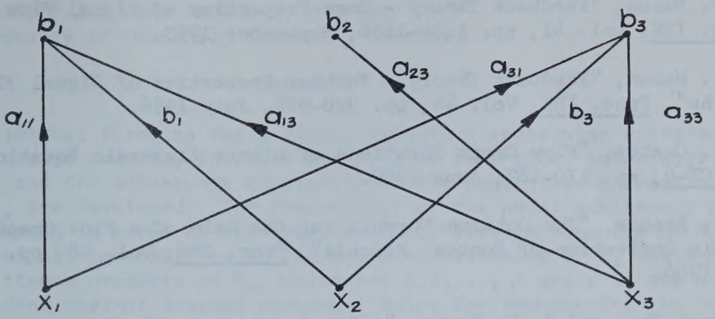


Fig. 8. The Graph G_2

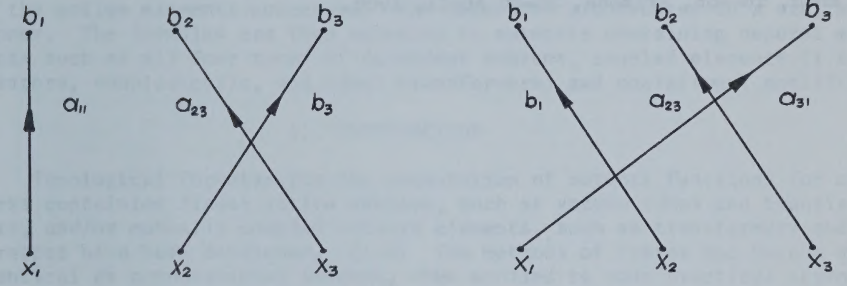


Fig. 9. The set of all the separations of G_2

Hence, in light of the theorem,

$$x_2 = \frac{\Delta_2}{\Delta} = \frac{(-1)^{2-1}\, a_{11}\, a_{23}\, b_3 + (-1)^{3-1}\, a_{23}\, a_{31}\, b_1}{+ a_{11}\, a_{22}\, a_{33} + (-1)^{3-1}\, a_{12} a_{23}\, a_{31} + (-1)^{2-1}\, a_{13}\, a_{22}\, a_{31}}$$

or,

$$x_2 = \frac{a_{23}\, a_{31}\, b_1 - a_{11}\, a_{23}\, b_3}{a_{11}\, a_{22}\, a_{33} + a_{12}\, a_{23}\, a_{31} - a_{13}\, a_{22}\, a_{31}}$$

IV. CONCLUSION

In this paper we have presented a new approach of solving a linear system of equations. We have also shown the simple derivation of the gain formula. The newly defined graphs are easy to construct since there exists symmetry in these graphs. Consequently, the set of all the separations can be determined with little effort.

When the system under study involves a large number of unknowns, the method of Chan and Bapna of determining the correct number of terms in the numerator and in the denominator of the gain formula can very well be applied here since there exists a one-to-one correspondence between the separations of G and the loop-sets and between the separations of G_j and the 2-loop-sets [5].

REFERENCES

[1] S. J. Mason, "Feedback Theory - Some Properties of Signal Flow Graphs", Proc. IRE, Vol. 41, pp. 1144-1156, September 1953.

[2] S. J. Mason, "Feedback Theory - Further Properties of Signal Flow Graphs", Proc. IRE, Vol. 44, pp. 920-926, July 1956.

[3] C. L. Coates, "Flow Graph Solutions of Linear Algebraic Equations", IRE CT-6, pp. 170-187, June 1959.

[4] C. A. Desoer, "The Optimum Formula for the Gain of a Flow Graph or a Simple Derivation of Coates' Formula", Proc. IRE, Vol. 48, pp. 883-889, May 1960.

[5] S. Park Chan and B. H. Bapna, "A New Approach to the Coates-Desoer Gain Formula", Convention Records, IEEE 1966 Region Six Annual Conference, Tucson, Arizona, 26-28 April, 1966.

TOPOLOGICAL FORMULAS FOR GENERAL LINEAR NETWORKS

MAW-TSUEY JONG AND GEORGE W. ZOBRIST
University of Missouri, Columbia, Missouri

ABSTRACT

Topological formulas for a linear network N containing voltage-controlled current sources, in which the passive reciprocal subnetwork is called N_p and the subnetwork consisting of the controlled current sources is called N_a, are developed. The determinant of the nodal admittance matrix of N is expressed in terms of the tree-branch admittance products of N_p and the products of all combinations of the k transadmittances of N_a and the (k+1)-tree admittance products of N_p, where $k=1,2,3,\ldots,\nu$ and ν is the number of the dependent current sources present. Rules for enumerating the required sets of (k+1)-trees of N_p are developed. All (k+1)-trees in the same set have the same sign which can be determined by considering the reduced graph of the active elements concerned. The cofactors are treated in a similar manner. The formulas are then extended to networks containing general elements such as all four types of dependent sources, coupled elements (i.e., gyrators, coupled-coils, and ideal transformers) and operational amplifiers.

I. INTRODUCTION

Topological formulas for the computation of network functions for networks containing linear active devices, such as vacuum tubes and transistors, and/or mutually coupled passive elements, such as transformers and gyrators have been developed. (1-6) The methods of Coates and Mayeda are identical as computational schemes, when applied to most practical networks. Their technique is to modify the graph in such a fashion that the resultant element-admittance matrix of the new graph is diagonal. Then they proceed to evaluate the network determinant and cofactors by finding all the complete-tree and complete 2-tree products of the new graph. The remaining problem is to find the sign for each term. In this procedure, each tree generated has to be checked for admissibility and some extra trees are generated, since only common trees of the voltage graph and current graph are used in the computation of the network determinant.

Based on the theories of Coates and Mayeda, this paper first expresses the determinant of the node-admittance matrix of a linear network N containing ν voltage-controlled current sources, in which the passive reciprocal subnetwork is called N_p and the subnetwork consisting of the controlled current sources is called N_a, in terms of the tree-branch admittance products of N_p and the products of all combinations of the k transadmittances of N_a and the (k+1)-tree admittance products of N_p, where $k=1,2,3,\ldots,\nu$. By considering the reduced graph of the active elements concerned, all (k+1)-trees in the same set have the same sign which can easily be determined. The cofactors are treated in a similar manner. For networks containing general linear elements such as all four types of dependent sources, coupled elements (i.e., gyrators, coupled-coils, and ideal transformers) and operational amplifiers, the subnetwork consisting of the active and/or nonreciprocal elements can be transformed equivalently into a subnetwork containing voltage-controlled current sources by employing some artifice and the simple source transformation. Then the developed topological formulas can be applied to obtain the desired network functions.

II. PRELIMINARY CONSIDERATIONS

We first state some theorems without proof which will be used in the following development.

A. Let G be a labelled linear connected graph. Let T be the sum of all tree products of G. Let $y_1, y_2, \ldots, y_n$ be the edges between vertex pairs $(i_1, j_1), (i_2, j_2), \ldots, (i_n, j_n)$. Then

$$\frac{\partial^k T}{\partial y_1 \partial y_2 \cdots \partial y_k} = \sum_{(V)} T_{V_1, V_2, \ldots, V_{k+1}}, \quad K = 1, 2, \cdots, n \tag{1}$$

where $T_{V_1, V_2, \ldots V_{k+1}}$ is the sum of (k+1)-tree products of G with the (k+1) subsets of vertices $V_1, V_2, \ldots, V_{k+1}$ in (k+1) separate parts, and (V) denotes that the summation is for all possible subsets of vertices formed in a manner shown in the following table for k=1,2,3.

K	$V_1, V_2, \cdots, V_{k+1}$
1	(i_1, j_1)
2	$(i_1 i_2, j_1, j_2)$ $(i_1 j_2, j_1, i_2)$ $(i_1, j_1 i_2, j_2)$ $(i_1, j_1 j_2, i_2)$
3	$(i_1 i_2 i_3, j_1, j_2, j_3)$ $(i_1 i_2 j_3, j_1, j_2, i_3)$ $(i_1 i_2, j_1 i_3, j_2, j_3)$ $(i_1 i_2, j_1 j_3, j_2, i_3)$
	$(i_1 i_2, j_1, j_2 i_3, j_3)$ $(i_1 i_2, j_1, j_2 j_3, i_3)$ $(i_1 j_2 i_3, j_1, i_2, j_3)$ $(i_1 j_2 j_3, j_1, i_2, i_3)$
	$(i_1 j_2, j_1 i_3, i_2, j_3)$ $(i_1 j_2, j_1 j_3, i_2, i_3)$ $(i_1 j_2, j_1, i_2 i_3, j_3)$ $(i_1 j_2, j_1, i_2 j_3, i_3)$
	$(i_1 i_3, j_1 i_2, j_2, j_3)$ $(i_1 j_3, j_1 i_2, j_2, i_3)$ $(i_1, j_1 i_2 i_3, j_2, j_3)$ $(i_1, j_1 i_2 j_3, j_2, i_3)$
	$(i_1, j_1 i_2, j_2 i_3, j_3)$ $(i_1, j_1 i_2, j_2 j_3, i_3)$ $(i_1 i_3, j_1 j_2, i_2, j_3)$ $(i_1 j_3, j_1 j_2, i_2, i_3)$
	$(i_1, j_1 j_2 i_3, i_2, j_3)$ $(i_1, j_1 j_2 j_3, i_2, i_3)$ $(i_1, j_1 j_2, i_2 i_3, j_3)$ $(i_1, j_1 j_2, i_2 j_3, i_3)$

It is not hard to see how each set for any k is formed from the previous set for (k-1). It can be shown that the number of subsets of vertices for each k is $2^{(k-1)} k!$.

The terms in T containing the products $y_1 y_2 \ldots y_k$ is given by

$$(y_1 y_2 \ldots y_k) \frac{\partial^k T}{\partial y_1 \partial y_2 \ldots \partial y_k}.$$

B. Let $V_1, V_2, \ldots V_n$ and $U_1, U_2, \ldots U_n$ be disjoint subsets of vertices of G. Then

$$T_{V_1, V_2, \cdots, V_n} \cap T_{U_1, U_2, \cdots, U_n} = \sum_{p(i)} T_{V_1 \cup U_{i_1}, V_2 \cup U_{i_2}, \cdots, V_n \cup U_{i_n}} \tag{2}$$

where p(i) denotes the permutations $i_1 i_2 \ldots i_n$ of $123 \ldots n$ and the summation is for all permutations $i_1 i_2 \ldots i_n$.

C. $T_{V_1, V_2, \ldots, V_n} = 0$ if $V_{i_1} \cap V_{i_2} \neq \emptyset$ for some $i_1, i_2 \in \{1, 2, \cdots, n\}$.

III. NETWORKS CONTAINING VOLTAGE-CONTROLLED CURRENT SOURCES

Consider a linear network N containing ν voltage-controlled current sources, in which the passive reciprocal subnetwork is called N_p and the subnetwork consisting of the controlled current sources is called N_a, as shown in Fig. 1-a. The voltage graph G_V and the current graph G_I of N are shown in Fig. 1-b & c, respectively. The determinant of the nodal admittance matrix, which is the sum of the common-tree admittance products of G_V and G_I, can be expressed as

$$\Delta_n = T = T^0 + \sum_{r=1}^{\nu} y_{m_r} T^r + \sum_{r \neq s}^{\nu} y_{m_r} y_{m_s} T^{rs} + \cdots + y_{m_1} y_{m_2} \cdots y_{m_\nu} T^{12\cdots\nu} \qquad (3)$$

where T^0 is the sum of all tree admittance products of N_p, and $T^{r_1 r_2 \cdots r_k}$ is the sum of (k+1)-tree admittance products of N_p which are found by the method explained in section II-A and II-B. If we consider the sign of each term in the sum as affixed to the (k+1)-tree product part of that term, then all (k+1)-trees in the same $T_{v_1, v_2, \ldots, v_{k+1}}$ have the same sign which can be determined by considering the reduced graph of the active elements concerned. For example, $T' = T_{i_1, j_1} \cap T_{k_1, \ell_1} = T_{i_1\ell_1, j_1k_1} - T_{i_1k_1, j_1\ell_1}$; all terms in $T_{i_1 l_1, j_1 k_1}$ have positive sign and all terms in $T_{i_1 k_1, j_1 l_1}$ are negative.

For practical network elements, such as vacuum tubes and transistors, the vertices i_r, j_r, k_r and l_r are not all distinct. Thus by section II-C, some sets of k-trees are empty and the labor to find all sets of k-trees is much reduced.

IV. GENERAL LINEAR NETWORKS

The above result can be extended to networks containing general linear elements such as all four types of dependent sources, coupled elements (i.e., gyrators, coupled-coils, and ideal transformers) and operational amplifiers. The technique is to transform the active and/or nonreciprocal elements equivalently into voltage-controlled current sources. This is shown in the following for each individual case.

A network containing a voltage-controlled voltage source is shown in Fig. 2-a. We can insert an impedance Z_0 in series with the voltage source and take the limit as $Z_0 \to 0$, as shown in Fig. 2-b. After making the simple source transformation, the voltage source is transformed into a current source with $Y_0 \to \infty$ as Fig. 2-c shows. Thus the case becomes one of voltage-controlled current source and can be solved by the previous method. Since we are interested in the network function, which is the ratio of the determinant and a cofactor, we can take only those terms containing Y_0 in T with Y_0 omitted for Δ_n,

$$\Delta_n = T_{k,\ell} + \mu\,(T_{i\ell,jk} - T_{ik,j\ell}) \qquad (4)$$

Example 1.

For the network shown in Fig. 3, we have

$$T_{1,2'} = y_1y_2 + y_1y_3 + y_1y_4 + y_2y_3 + y_2y_5 + y_3y_4 + y_3y_5 + y_4y_5$$

$$T_{12,1'2'} = y_2y_5 \qquad\qquad T_{11',22'} = y_1y_4$$

$$\Delta_n = T_{1,2'} + \mu\,(T_{11',22'} - T_{12,1'2'})$$

$$= (1+\mu)y_1y_4 + (1-\mu)y_2y_5 + y_1y_2 + y_1y_3 + y_2y_3 + y_3y_4 + y_3y_5 + y_4y_5$$

$$\Delta_{11'} = (1+\mu)y_4 + y_2 + y_3$$

$$Y_{11'} = \frac{\Delta_n}{\Delta_{11'}} = y_1 + \frac{(1-\mu)y_2y_5 + y_2y_3 + y_3y_4 + y_3y_5 + y_4y_5}{(1+\mu)y_4 + y_2 + y_3}$$

For the current-controlled voltage source in Fig. 4-a, if the controlling current flows through a finite admittance y_{ij}, it is essentially a voltage-controlled voltage source with voltage gain $\mu = z_m y_{ij}$. If the current is in a short-circuit, $y_{ij} = \infty$ and we take terms containing y_{ij} for Δ_n

with y_{ij} omitted,

$$\Delta_n = T_{ik,j,\ell} + T_{i\ell,j,k} + T_{i,jk,\ell} + T_{i,j\ell,k} + z_m (T_{i\ell,jk} - T_{ik,j\ell}) \tag{5}$$

Similarly, the current-controlled current source in Fig. 4-b with finite y_{ij} is essentially a voltage-controlled current source with transadmittance $y_m = \alpha y_{ij}$. Again, if $y_{ij} = \infty$,

$$\Delta_n = T_{i,j} + \alpha (T_{i\ell,jk} - T_{ik,j\ell}) \tag{6}$$

A common-emitter transistor is an example of a current-controlled current source in which the controlling base-current flows through the base admittance which can be considered as a short-circuit.

A gyrator can be considered as consisting of two voltage-controlled current sources, as shown in Fig. 5. Hence applying the formula, we have

$$\Delta_n = T^0 + (y_{g_2} - y_{g_1}) T^1 + y_{g_1} y_{g_2} T^{12} \tag{7}$$

where

$$\begin{aligned} T^0 &= \text{sum of tree products of } N_p \\ T^1 &= T_{i\ell,jk} - T_{ik,j\ell} \\ T^{12} &= T_{ik,j,\ell} + T_{i\ell,j,k} + T_{i,jk,\ell} + T_{i,j\ell,k} \end{aligned} \tag{8}$$

If $y_{g_1} = y_{g_2} = y_g$, $\quad \Delta_n = T^0 + y_g^2 T^{12}$ (9)

The coupled coils (a finite-coupling transformer) in Fig. 6-a is equivalent to a gyrator with the two ports shunted by two branches as shown in Fig. 6-b. After some manipulations we obtain

$$\begin{aligned} \Delta_n = T^0 + \tfrac{1}{s^2\Delta} T^{12} + \tfrac{1}{s\Delta}\big[&(L_1 + L_2 + 2M) T_{i\ell,jk} + (L_1 + L_2 - 2M) T_{ik,j\ell} \\ &+ L_1 (T_{k,ij\ell} + T_{\ell,ijk}) + L_2 (T_{i,jk\ell} + T_{j,ik\ell})\big] \end{aligned} \tag{10}$$

where T^0 and T^{12} are given by Eq. 8, and $\Delta = L_1 L_2 - M^2$.

Example 2. Consider the network in Fig. 6-c.

$$T^0 = y_3 y_4 y_5 \qquad T^{12} = y_3 + y_4 + y_5 \qquad T_{12',1'2} = 0 \qquad T_{12,1'2'} = y_3 y_4$$

$$T_{1,1'22'} = 0 \qquad T_{1',122'} = y_3 y_5 \qquad T_{2,11'2'} = y_4 y_5 \qquad T_{2',11'2} = 0$$

$$\Delta_n = y_3 y_4 y_5 + \tfrac{1}{s^2\Delta} (y_3 + y_4 + y_5) + \tfrac{1}{s\Delta}\left[(L_1 + L_2 - 2M) y_3 y_4 + L_1 y_4 y_5 + L_2 y_3 y_5\right]$$

In this example, the voltage graph and current graph each have 30 trees, but only the 9 common trees are used in the computation of Δ_n. It is seen how labor is saved by the k-tree method.

For a perfect transformer, k=1, $M^2 = L_1 L_2$, $\Delta = 0$. So we take terms containing Δ only. With $\frac{1}{s\Delta}$ omitted, we have

$$\begin{aligned} \Delta_n = \tfrac{1}{s} T^{12} &+ (L_1 + L_2 + 2M) T_{i\ell,jk} + (L_1 + L_2 - 2M) T_{ik,j\ell} \\ &+ L_1 (T_{k,ij\ell} + T_{\ell,ijk}) + L_2 (T_{i,jk\ell} + T_{j,ik\ell}) \end{aligned} \tag{11}$$

An ideal transformer can be represented by a perfect transformer shunted in the primary winding by an inductance equal to the negative of the self-inductance of the primary winding. This is shown in Fig. 7. Thus for a network containing an ideal transformer, the nodal determinant is computed as

$$\begin{aligned} \Delta_n = (1+a)^2 T_{i\ell,jk} + (1-a)^2 T_{ik,j\ell} &+ a^2 (T_{i,jk\ell} + T_{j,ik\ell}) \\ &+ T_{k,ij\ell} + T_{\ell,ijk} \end{aligned} \tag{12}$$

Example 3. In Fig. 6-c, if the practical transformer is replaced by an ideal transformer, we have

$$\Delta_n = (1-a)^2 y_3 y_4 + a^2 y_3 y_5 + y_4 y_5 \qquad \Delta_{11'} = y_3 + y_4 + y_5$$

$$Y_{11'} = \frac{\Delta_n}{\Delta_{11'}} = \frac{(1-a)^2 y_3 y_4 + a^2 y_3 y_5 + y_4 y_5}{y_3 + y_4 + y_5}$$

which agrees with Talbot.(4)

V. CONCLUSIONS

General linear networks can be analyzed by applying the topological formulas developed in this paper. By dividing the vertices of the active and/or nonreciprocal elements into the required subsets and generating the corresponding k-trees of the passive reciprocal subnetwork, the network functions can be found without modifying the network topology. With the determinant and cofactors expressed in the form developed, the effect of an active and/or nonreciprocal element can be ascertained without calculating the entire network function. This fact would seem to be quite advantageous in the study of the sensitivity problem. The problem remaining is the generation of k-trees of a linear graph with specified subsets of vertices in separate parts without first generating all trees of the graph, as was done by Hakimi and Green.(7)

ACKNOWLEDGMENTS

This paper was supported by the University of Missouri Engineering Experiment Station. This paper also represents a portion of the work in progress on a Ph.D. dissertation, by the first author, at the University of Missouri, Columbia, Missouri.

VI. REFERENCES

1. Mason, S. J., "Topological Analysis of Linear Non-reciprocal Networks," Proc. IRE, Vol. 45, pp. 829-838; June, 1957.

2. Coates, C. L., "General Topological Formulas for Linear Network Functions," IRE Trans. CT-5, pp. 30-42; March, 1958.

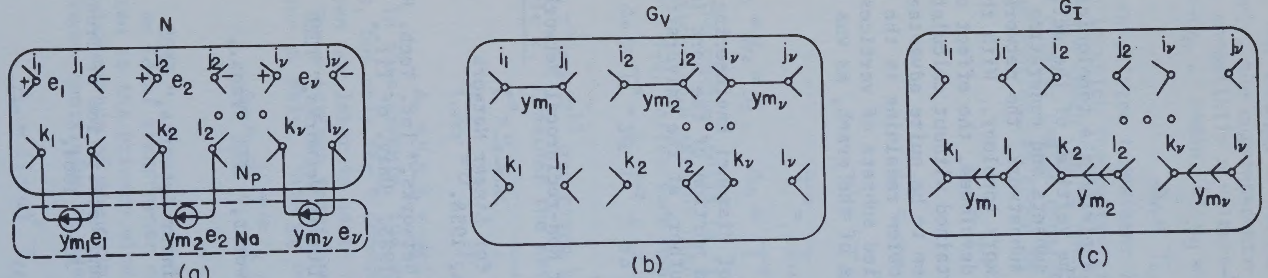
3. Mayeda, W., "Topological Formulas for Active Networks," Int. Tech. Rept., No. 8, U. S. Army Contract No. DA-11-022-ORD-1983. Univ. of Ill., Urbana, Ill.; Jan. 1958.

4. Talbot, A., "Topological Analysis of General Linear Networks," IEEE Trans., CT-12, pp. 170-180; June, 1965.

5. Nathan, A., "Topological Rules for Linear Networks," IEEE Trans., CT-12, pp. 344-358; Sept. 1965.

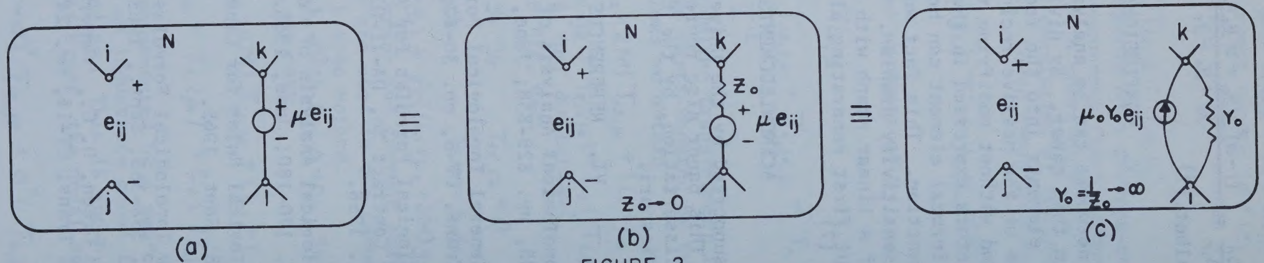
6. Brown, D. P., "New Topological Formulas for Linear Networks," IEEE Trans., CT-12, pp. 358-365; Sept., 1965.

7. Hakimi, S. L., and Green, D. G., "Generation and Realization of Trees and k-Trees," IEEE Trans. CT-11, pp. 247-255; June, 1964.


N
G_V
G_I
i_1 j_1 i_2 j_2 i_ν j_ν
e_1 e_2 e_ν
k_1 l_1 k_2 l_2 k_ν l_ν
N_p
$y_{m_1}e_1$ $y_{m_2}e_2$ Na $y_{m_\nu}e_\nu$
y_{m_1} y_{m_2} y_{m_ν}

(a) (b) (c)

FIGURE 1


N
i
j
k
l
e_{ij}
μe_{ij}
≡
Z_o
$Z_o \to 0$
$\mu_o Y_o e_{ij}$
Y_o
$Y_o = \frac{1}{Z_o} \to \infty$

(a) (b) (c)

FIGURE 2

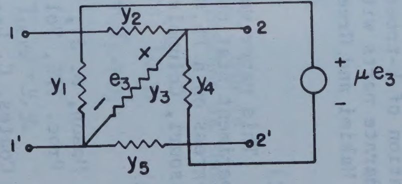

1
1'
2
2'
y_1
y_2
y_3
y_4
y_5
e_3
x
μe_3

FIGURE 3

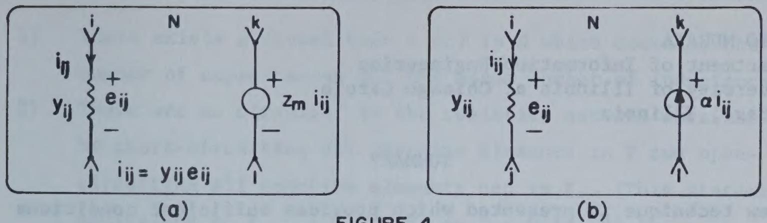

N
i
j
i_{ij}
y_{ij}
e_{ij}
+
−
k
l
$z_m\, i_{ij}$
$i_{ij} = y_{ij} e_{ij}$
(a)
N
i
j
i_{ij}
y_{ij}
e_{ij}
k
l
αi_{ij}
(b)

FIGURE 4

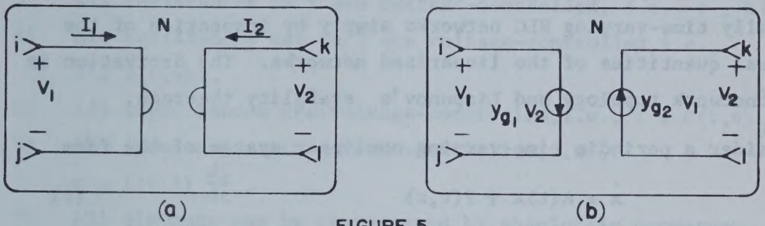

N
I_1
I_2
i
j
k
l
V_1
V_2
+
−
(a)
N
i
j
k
l
V_1
V_2
$y_{g_1} V_2$
$y_{g_2} V_1$
(b)

FIGURE 5

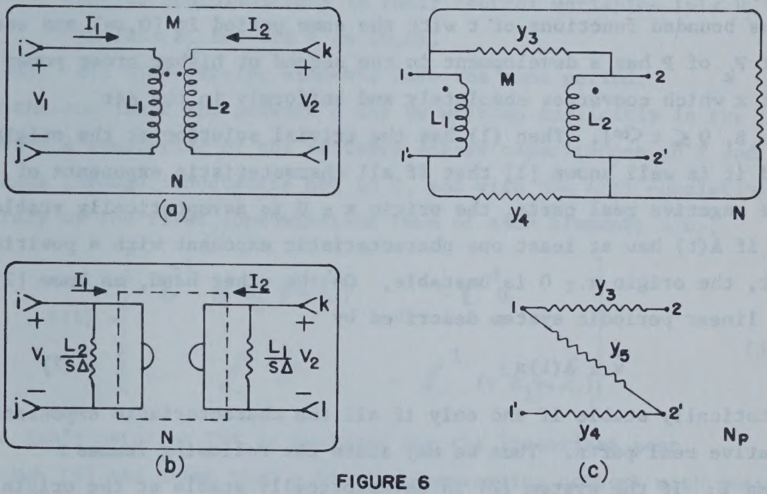
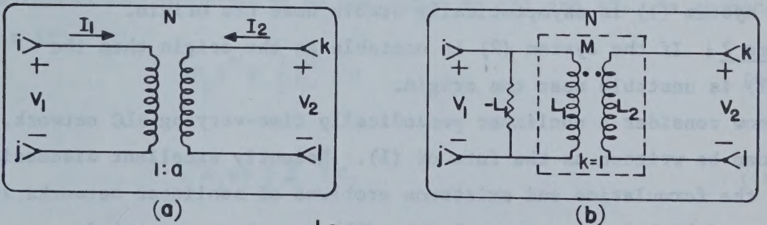

M
I_1
I_2
i
j
k
l
V_1
V_2
L_1
L_2
N
(a)
y_3
1
2
1'
2'
M
L_1
L_2
y_5
y_4
N
I_1
I_2
i
j
k
l
V_1
V_2
$\frac{L_2}{s\Delta}$
$\frac{L_1}{s\Delta}$
N
(b)
y_3
1
2
y_5
1'
2'
y_4
N_P
(c)

FIGURE 6


N
I_1
I_2
i
j
k
l
V_1
V_2
1:a
(a)
N
M
i
j
k
l
V_1
V_2
$-L_1$
L_1
L_2
k=1
(b)

$$a^2 = \frac{L_2}{L_1}, \; L_1 = 1, \; L_2 = a^2, \; M = a$$

FIGURE 7

NETWORK TOPOLOGY AND STABILITY OF NONLINEAR RLC NETWORKS*

TADAO MURATA
Department of Information Engineering
University of Illinois at Chicago Circle
Chicago, Illinois

SUMMARY

A new technique is presented which provides sufficient conditions for local stability and instability of a certain class of nonlinear periodically time-varying RLC networks simply by inspection of the topological quantities of the linearized networks. The derivation is based on network topology and Liapunov's stability theorems.

Consider a periodic time-varying nonlinear system of the form

$$\dot{x} = A(t)x + P(t,x) \tag{1}$$

where A(t) is a real square matrix whose elements are constants or continuous bounded functions of t with the same period in $[0,\infty)$ and each component P_k of P has a development in the second or higher order power series of x which converges absolutely and uniformly in the set $S = [\,\|x\| < B,\ 0 \leq t < \infty]$. Then (1) has the trivial solution at the origin $x = 0$ and it is well known [1] that if all characteristic exponents of A(t) have negative real parts, the origin $x = 0$ is asymptotically stable and that if A(t) has at least one characteristic exponent with a positive real part, the origin $x = 0$ is unstable, On the other hand, we know [2] that the linear periodic system described by

$$\dot{x} = A(t)x \tag{2}$$

is asymptotically stable if and only if all the characteristic exponents have negative real parts. Thus we may state the following lemmas:

Lemma 1: If the system (2) is asymptotically stable at the origin then the system (1) is asymptotically stable near the origin.

Lemma 2: If the system (2) is unstable at the origin then the system (1) is unstable near the origin.

We now consider a nonlinear periodically time-varying RLC network, N which can be written as the form of (1). Recently excellent discussions [3-7] on the formulation and existence problems of nonlinear networks in the normal form (which is more general than (1)) have been presented.

---

*This work was supported in part by the Joint Services Electronics Program under Contract No. DA 28 AMC 00073(E) and in part by the Air Force Office of Scientific Research under Grant No. AF AFOSR 931.65.

Let the following conditions be satisfied:

1) There exists a normal tree T [8] in N which contains the maximum number of capacitances and the minum number of inductances.
2) There are no circuits[1] in the resistive network obtained from N by short-circuiting all reactive elements in T and open-circuiting all reactive elements not in T. (This statement is equivalent to $F_{RG} = 0$ in [8] or $\delta = 0$ in [9].)
3) All resistances in T are current-controlled, i.e., $v = R(t,i)i$.
4) All resistances not in T are voltage-controlled, i.e., $i = G(t,v)v$.
5) All capacitances are voltage-controlled, i.e., $i = C(t,v)\frac{dv}{dt}$.
6) All inductances are current-controlled, i.e., $v = L(t,i)\frac{di}{dt}$.
7) All elements can be represented by absolutely convergent power series approximations in their control variables $|v| < B_0$ or $|i| < B_0$ for all t in $[0,\infty)$.
8) All time-varying elements have the same period.

Then the nonlinear RLC network N may be written explicitly in the form of (1) with x consisting of the voltages across capacitances in T and the currents through inductances not in T, and with the A(t) consisting entirely of the first approximation term of each element, i.e.,

$$A(t) = \begin{bmatrix} -\mathcal{C}^{-1}(\beta G_2 \beta^t + \dot{\mathcal{C}}) & -\mathcal{C}^{-1}\alpha \\ \mathcal{L}^{-1}\alpha^t & -\mathcal{L}^{-1}(\gamma^t R_1 \gamma + \dot{\mathcal{L}}) \end{bmatrix} \qquad (3)$$

where Kuh's notation [9] is employed for the linearized part.

Kuh [9] has shown that a linear time-varying network with positive capacitances and inductances as described by (2) and (3), is asymptotically stable if

$$\beta G_2 \beta^t + \frac{1}{2}\dot{\mathcal{C}} > 0 \qquad (4a)$$

and

$$\gamma^t R_1 \gamma + \frac{1}{2}\dot{\mathcal{L}} > 0. \qquad (4b)$$

---

[1] A self-loop is not considered to be a circuit here.

Similarly using the instability theorem [10] it can be shown that the above network is unstable if

$$\beta G_2 \beta^t + \frac{1}{2}\dot{\mathbb{C}} < 0 \tag{5a}$$

$$\gamma^t R_1 \gamma + \frac{1}{2}\dot{\mathcal{L}} < 0 \; . \tag{5b}$$

Thus from Lemmas 1 and 2, we may state the following theorem:

Theorem 1: A periodic time-varying nonlinear network with positive first approximation terms of capacitances and inductances as described by (1) and (3), is asymptotically stable near the origin if the condition (4) holds and unstable near the origin if the condition (5) holds.

Recently the author [11] has obtained topological formulas[2] for the determinants and principal minors of the matrices in a more general form than those of (4) by interpreting them as short-circuit and open-circuit operators of the special resistive multiport networks which include the time-derivatives of reactive elements, $\dot{C}$ and $\dot{L}$. For example, the principal minors of order r of the matrix $\beta G_2 \beta^t + \frac{1}{2}\dot{\mathbb{C}}$ of order n are given in terms of the sum of tree products of (the first approximation terms of) conductances in the time-domain as:

$$y_r = \frac{\Sigma(\text{tree products for the network } N_{C/2} + n-4)}{\Sigma(\text{tree products for the network } N_{C/2} + n)} \tag{6}$$

$$r = n, n-1, \ldots, 1 \; .$$

The networks $N_{C/2} + n-r$ and $N_C + n$ are those obtained from $N_{C/2}$ by shorting (n-r) ports and all the n ports, respectively, where $N_{C/2}$ is the resistive multiport faced by the capacitances in T when one half of its time-derivative, $\frac{1}{2}\dot{C}_i$ is connected in parallel with every time varying capacitance, C_i and all the remaining reactive elements are open-circuited. Similarly, the principal minors of order r of the matrix $\gamma^r R_1 \gamma + \frac{1}{2}\dot{\mathcal{L}}$ of order m are given in terms of the sum of chord-set products of (the first approximation terms of) resistances in the time-domain as

$$z_r = \frac{\Sigma(\text{chord-set products for the network } N_{L/2} + m)}{\Sigma(\text{chord-set products for the network } N_{L/2} + m-r)} \tag{7}$$

$$r = m, m-1, \ldots, 1 \; .$$

---

[2] Mutual inductances are not considered for simplicity of topological formulas.

The networks $N_{L/2} + m$ and $N_{L/2} + m-r$ are those obtained from $N_{L/2}$ by adding source-branches at all the m ports and at (m-r) ports, respectively, where $N_{L/2}$ is the resistive multiport faced by the inductances not in T when one half of its time-derivative, $\frac{1}{2}\dot{L}_i$ is connected in series with every time-varying inductances L_i and all the remaining reactive elements are short-circuited.

We are now in a position to interpret the conditions in Theorem 1 in terms of the topological formulas (6) and (7) and state the following theorem:

Theorem 2 A periodic time-varying nonlinear network with positive first approximation terms of capacitances and inductances and with no mutual inductances as described by (1) and (3), is asympotically stable near the origin if $y_r > 0$ for $r = n,n-1,\ldots,1$ and $z_r > 0$ for $r = m,m-1,\ldots,1$. It is unstable if $y_r < 0$ for $r = n,n-1,\ldots,1$ and $z_r < 0$ for $r = m,m-1,\ldots,1$.

The use of Theorem 2 makes it possible to obtain stability information or stability criteria simply by inspection of the topology and elements of the linearized networks of a nonlinear network without solving or formulating network equations.

The author is grateful to Professors M.E. Van Valkenburg, W. Mayeda and J.B. Cruz, Jr. for valuable advice and his colleague, Dr. G.O. Martens for many helpful remarks.

References

1. E.A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.

2. L.A. Zadeh and C.A. Desoer, Linear System Theory, McGraw-Hill, New York, 1963.

3. R.K. Brayton and J.K. Moser, "A Theory of Nonlinear Networks," J. Appl. Math., Vol. XXII, April and July 1964, pp. 1-33 and 81-104.

4. L.O. Chua and R.A. Rohrer, "On the Dynamic Equations of a Class of Nonlinear RLC Networks," IEEE Trans. on Circuit Theory, Vol. CT-12, December, 1965, pp. 475-489.

5. C.A. Desoer and J. Katzenelson, "Nonlinear RLC Networks," Bell Syst. Tech.J., Vol. 44, January, 1965, pp. 161-198.

6. R. Liu and L.V. Auth, "Qualitative Synthesis of Nonlinear Networks," Proc. First Annual Allerton Conference on Circuit and System Theory, November, 1963, pp. 330-343.

7. T.E. Stern, "On Equations of Nonlinear Networks, "IEEE Trans. on Circuit Theory, Vol. CT-13, March, 1966, pp. 74-81.

8. E.S. Kuh and R.A. Rohrer, "The State-Variable Approach to Network Analysis, "Proc. IEEE, July, 1965, pp. 672-686.

9. E.S. Kuh, "Stability of Linear Time-Varying Networks - The State-Space Approach, "IEEE Trans. on Circuit Theory, June, 1965, pp. 150-157.

10. G. LaSalle and S. Lefschetz, Stability by Liapunov's Direct Method with Applications, Academic Press, New York, 1961.

11. T. Murata, "A Topological Approach to the Stability of Time-Varying and Nonlinear Networks, "Ph.D. Dissertation, University of Illinois, Urbana, Ill. (CSL Report R-291, May 1966), also "A Topological Approach to the Stability of Linear Time-Varying Networks," submitted for publication.

DIGITAL COMPUTER SYNTHESIS OF ADMITTANCE MATRICES OF N+1 NODES

ELMER A. HOYER
Department of Electrical Engineering, Wichita State University, Wichita, Kansas
GEORGE PERDIKARIS
General Electric Co., Bloomington, Illinois
GEORGE W. ZOBRIST
Department of Electrical Engineering, University of Missouri, Columbia, Missouri

ABSTRACT

The problem of realizing a Y-matrix with an N-port network of N+1 nodes has been widely discussed and solved by various authors. There are two major approaches to the synthesis of these networks. The first approach is to realize the port-structure through the sign pattern of the Y-matrix. The second approach is based on the decomposition of the Y-matrix into the triple product of $C_s \, Y_e \, C_s^t$ where C_s is a seg matrix and Y_e is a diagonal matrix of positive elements.

The approach of this paper is to combine some of the best qualities of both methods, for a Y-matrix with all non-zero elements. A computer program is developed which decomposes the Y-matrix, which is a function of the complex S, using Cederbaum's algorithm. This is accomplished by letting S = 1 and decomposing the conductance matrix so formed. A set of simultaneous linear equations are then developed, for which the coefficient matrix is a triangular matrix, to solve for the diagonal elements of Y_e.

The port structure of the network is determined through the sign pattern of the Y-matrix. The computer program develops a matrix which gives the ports connected to each of the N+1 nodes and develops another matrix which gives the nodes to which each port is connected. From either of these matrices the port structure can be developed. Once the port structure is known the synthesis is completed by the use of the cut-set matrix and the element matrix Y_e.

FREQUENCY SENSITIVITY OF ACTIVE NETWORKS

E. C. BALL
Lieutenant Commander
Royal Canadian Navy

G. J. THALER
U. S. Naval Postgraduate School
Monterey, California

ABSTRACT: Sensitivity analysis is applied to the frequency stability of RC Oscillators with respect to variation in system parameters. Analytical sensitivity functions are determined from the characteristic equations of the oscillators. Comparison of parameter plane curves and sensitivities for several structures are used as a basis for oscillator design to meet frequency and frequency sensitivity specifications.

INTRODUCTION

The stability of frequency in oscillators has been studied from many points of view. The parameter plane methods of analyzing self-excited oscillation in non-linear systems, suggested by Siljak[1] and Siljak and Stojic[2] lend themselves to a new visualization of electronic oscillators.

In this paper, the suggested analysis is applied to several common active networks and a number of computational short-cuts are pointed out for dealing with the second and third order systems which arise.

The Horowitz[3] definition of sensitivity is used throughout, with notation as follows:

$$S_{\alpha}^{\omega} = \frac{\frac{\partial\omega}{\omega}}{\frac{\partial\alpha}{\alpha}} = \frac{\alpha}{\omega} \cdot \frac{\partial\omega}{\partial\alpha} = \frac{\partial \ln \omega}{\partial \ln \alpha}$$

Features of mathematical tractibility and normalization inherent in this definition and the specification of tolerances on components as a percentage make this form useful in this application.

FREQUENCY SENSITIVITY OF OSCILLATORS

In oscillators, amplitude stabilization is achieved by the onset of a non-linear property of some element, often the active element of the network. The most common example of this is saturation in an amplifier. This leads to the establishment of a stable limit cycle. In this paper, the non-linear active element is a device whose voltage transfer is a simple gain (positive or negative) which is a function only of the amplitude, and which introduces no phase change. This limits the scope of this analysis since it is known that frequency stability of oscillators depends on changes in network parameters, phase shift in amplifiers, changes in input and output impedance and changes in harmonic content of the signal. In many practical variable frequency oscillators, the circuit is compensated to make the amplitude relatively insensitive to change of frequency. The effects of these adaptations are not considered.

The non-linear gain is chosen as one of the variables in the parameter plane. The other variable is assigned to a parameter of interest. The values of all other parameters are fixed at some reasonable value. The coefficients of the characteristic equation are manipulated in the usual parameter plane method to find expressions for the two parameters. In this plane, the non-linear gain gives rise to an M-locus which is a function of the amplitude. By examining the parameter plane and the M-locus, the existence of a limit cycle may be observed, its stability determined, and from the values of the parameters at the intersection of the $\zeta = 0$ curve and the M-locus, the frequency of the limit cycle may be determined.

If the limit cycle is stable, it is seen that a change in the network parameter of interest will cause the steady state condition to change. In effect, the steady state M-point moves along the $\zeta = 0$ locus in the parameter plane. Since the steady state M-point is characterized by the frequency and amplitude of the limit cycle, the slope of the $\zeta = 0$ curve must be directly related to the frequency and amplitude sensitivity. In accordance with the chain rule

$$\frac{\partial\beta}{\partial\alpha} = \frac{\partial\beta}{\partial\omega} \cdot \frac{\partial\omega}{\partial\alpha}, \text{ it is seen that } S_{\alpha}^{\beta} = S_{\omega}^{\beta} \cdot S_{\alpha}^{\omega}$$

If β is the parameter assigned to the variable gain, the sensitivity of frequency with respect to α, S_{α}^{ω}, may be determined from the parametric equation of the $\zeta = 0$ curve, $\alpha(\omega) = 0$, or it may be determined from the slope of the $\zeta = 0$ curve if there is concomitant knowledge of S_{ω}^{β}.

Of primary importance to the analysis is the derivation of the parametric equations for the $\zeta = 0$ contour in the parameter plane. If the analysis is to be done graphically, the $\zeta = 0$ contour must be constructed. To obtain the parametric equations of the $\zeta = 0$ curve, the methods of Mitrovic[4] or Siljak[5] may be used, but it is also possible to substitute $s=j\omega$ in the characteristic equation, separate the real and imaginary parts, set them equal to zero and solve the two simultaneous equations for the parameters α and β as functions of ω.

Since the parametric equations of the $\zeta = 0$ curve are ultimately determined from the characteristic equation, it will be of some utility to determine a relationship between frequency sensitivity with respect to a parameter variation and the coefficients of the characteristic equation.

For the general second order system,

$$a_2 s^2 + a_1 s + a_o = 0$$

Substituting $s = j\omega$, it is seen that

$$\omega^2 = \frac{a_o}{a_2} \quad \text{and} \quad a_1 = 0$$

This gives the characteristic frequency and conditions for oscillation (two roots on the $j\omega$ axis). The frequency sensitivity with respect to any parameter x, S_x^{ω}, is then determined from

$$2 \ln \omega = \ln a_o - \ln a_2$$

$$S_x^{\omega} = \frac{1}{2} (S_x^{a_o} - S_x^{a_2})$$

For a third order system,

$$a_3 s^3 + a_2 s^2 + a_1 s + a_o = 0$$

from which it is seen that

$$S_x^{\omega} = \frac{1}{2} (S_x^{a_1} - S_x^{a_3}) = \frac{1}{2} (S_x^{a_o} - S_x^{a_2})$$

It must be borne in mind that the basis of this system is the argument that there is a self-adjusting parameter in the system causing the steady state M-point to follow the $\zeta = 0$ curve. Since this parameter will be included in some or all of the coefficients of the characteristic equation, the partial derivative (or sensitivity) of a coefficient with respect to a parameter must also include the partial derivative of the adaptive parameter in accordance with the chain rule.

APPLICATION TO RC OSCILLATORS

Comparison of the parameter plane curves and sensitivity functions of several forms of oscillators yields comparative information on feasibility, required amplification, choice of configurations and parameter values to meet a priori frequency and sensitivity specifications. It is useful to compare several forms of RC oscillators on this basis. In these oscillators, use is made of the frequency selective or phase shift characteristics of RC networks to provide the frequency stabilization. In general, the feedback oscillator can be represented by the block diagram in figure 1. The general characteristic equation is then

$$1 - \beta H(s) = 0$$

The amplifier is assumed to be a perfect voltage amplifier with saturation.

The problem for this analysis is assumed to be to find a configuration and determine network parameters to minimize the sensitivity of the frequency of an RC oscillator to variations in a terminal resistance. The variable parameter might be the input impedance of the active device in the network or of a load. The parameter plane curves are drawn for a normalized frequency $x = RC\omega$, thus making the analysis applicable to any frequency range. To provide a particular basis for the comparison, all the resistances other than the parameter of interest are chosen equal to R and all capacitances equal to C. The parameter of interest is αR and is marked (*) in each figure. It is recognized that this may not lead to the best choice of parameters from other points of view. Mehta[6] has shown that this choice is not optimum to minimize sensitivity to phase change in the amplifier. However, this does not affect the method of which this is intended to be exemplary.

Six different configurations are considered. The first two are phase shift oscillators in which the amplifier provides a π shift and the network provides an additional π shift. The π shift of the amplifier is readily seen in the requirement for the gain, β, to be negative. The last four have frequency fixed by a phase crossover frequency. In each case the general characteristic equation of the oscillator is given and then the reduced characteristic equation for the particular example, where $p = RCs$. The parametric equations of the $\zeta = 0$ curve and the frequency sensitivity, S^x_α, are listed.

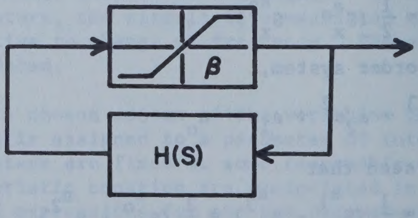


figure 1. Oscillator Block Diagram

Phase Shift Oscillator - High Pass

Characteristic Equation

$$s^3(1-\beta)R_1R_2R_3C_1C_2C_3+s^2(R_1R_2C_1C_2+R_2R_3C_2C_3+R_1R_3C_1C_3+R_1R_2C_1C_3+R_1R_2C_2C_3+R_1R_3C_2C_3)+s(R_1C_1+R_2C_2+R_3C_3+R_1C_2+R_2C_3)+1 = 0$$

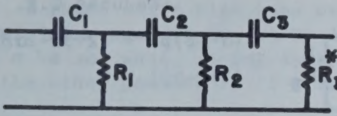


figure 2.

Reduced C. E.

$$(\alpha-\alpha\beta)p^3 + (3+3\alpha)p^2 + (4+\alpha)p + 1 = 0$$

$$\alpha = \frac{1-3x^2}{3x^2}, \qquad -\beta = \frac{12}{1-3x^2} + \frac{1}{x^2} - 1, \qquad S_\alpha^x = -\frac{\alpha}{2(1+\alpha)}$$

Phase Shift Oscillator - Low Pass

Characteristic Equation

$$s^3(R_1R_2R_3R_4C_1C_2C_3)+s^2(R_1R_2R_4C_1C_2+R_1R_2R_4C_1C_3+R_1R_2R_3C_1C_2+R_2R_3R_4C_2C_3+R_1R_3R_4C_1C_3+R_1R_3R_4C_2C_3)+s(R_1R_4C_1+R_2R_4C_2+R_3R_4C_3+R_2R_3C_2+R_1R_3C_1+R_1R_4C_3+R_2R_4C_3+R_1R_2C_1+R_1R_4C_2+R_1R_3C_2)+R_1+R_2+R_3+(1-\beta)R_4 = 0$$

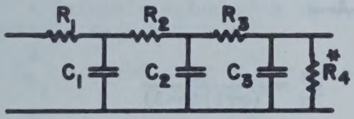


figure 3.

Reduced C. E.

$$\alpha p^3 + (5\alpha+1)p^2 + (6\alpha+4)p + (3+\alpha-\alpha\beta) = 0$$

$$\alpha = \frac{4}{x^2-6}, \qquad -\beta = \frac{x^4+11x^2+14}{4}, \qquad S_\alpha^x = \frac{-1}{3\alpha+2}$$

RC Oscillator - Wien Network

Characteristic Equation

$$s^2(R_1R_2C_1C_2)+s[R_1C_1+R_2C_2+(1-\beta)R_2C_1]+1 = 0$$

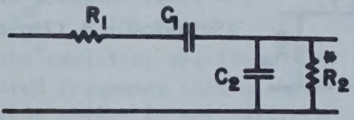


figure 4.

Reduced C. E.

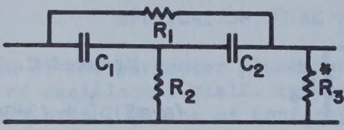
$$\alpha p^2 + (1+2\alpha-\alpha\beta)p + 1 = 0$$

$$\alpha = \frac{1}{x^2}, \qquad \beta = x^2 + 2, \qquad S_\alpha^x = -\frac{1}{2}$$

RC Oscillator - Bridged "T" Network - Type 1.

Characteristic Equation

$$s^2(1-\beta)R_1R_2R_3C_1C_2+s[(1-\beta)(R_2R_3C_1+R_2R_3C_2)+R_1R_3C_2+R_1R_2C_1+R_1R_2C_2]+R_1+(1-\beta)R_3 = 0$$



Reduced C.E.

$$(\alpha-\alpha\beta)p^2 + (2+3\alpha-2\alpha\beta)p + (1+\alpha-\alpha\beta) = 0$$

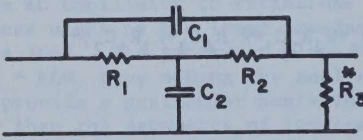
figure 5.

$$\alpha = \frac{2x^2}{1-x^2}, \qquad \beta = \frac{1+2x^2}{2x^2}, \qquad S^x_\alpha = \frac{1}{2+\alpha}$$

RC Oscillator - Bridged "T" Network - Type 2.

Characteristic Equation

$$s^2(1-\beta)R_1R_2R_3C_1C_2+s[(1-\beta)(R_1R_3C_1+R_2R_3C_1)+R_1R_3C_2+R_1R_2C_2]+R_1+R_2+(1-\beta)R_3 = 0$$



Reduced C.E.

$$(\alpha-\alpha\beta)p^2 + (1+3\alpha-2\alpha\beta)p + (2+\alpha-\alpha\beta) = 0$$

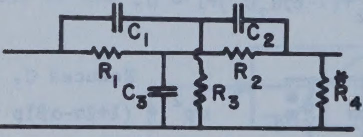
figure 6.

$$\alpha = \frac{3+x^2}{1-x^2}, \qquad \beta = \frac{5+x^2}{3+x^2}, \qquad S^x_\alpha = \frac{2\alpha}{(\alpha+1)(\alpha-3)}$$

RC Oscillator - Twin "T" Network

Characteristic Equation

$$s^3(1-\beta)R_1R_2R_3R_4C_1C_2C_3+s^2[(1-\beta)(R_1R_3R_4C_1C_2+R_2R_3R_4C_1C_2)+R_1R_2R_4C_2C_3+R_1R_2R_3C_1C_3+R_1R_2R_3C_2C_3+R_1R_3R_4C_2C_3+R_1R_3R_4C_1C_3]+s[(1-\beta)(R_3R_4C_1+R_3R_4C_2)+R_2R_3C_2+R_2R_3C_1+R_1R_3C_2+R_1R_3C_1+R_2R_4C_2+R_1R_4C_2+R_1R_4C_3+R_1R_2C_3]+[R_1+R_2+(1-\beta)R_4] = 0$$



Reduced C.E.

$$(\alpha-\alpha\beta)p^3+ (2+5\alpha-2\alpha\beta)p^2 + (5+5\alpha-2\alpha\beta)p + (2+\alpha-\alpha\beta) = 0$$

figure 7.

$$\alpha = \frac{2x^4+4x^2-1}{3-3x^4}, \qquad \beta = \frac{2x^6+9x^4-21x^2-10}{2x^6-9x^2+2}, \qquad S^x_\alpha = \tfrac{1}{2}[S^\alpha_{x^2}]^{-1}$$

where $$S^\alpha_{x^2} = -\frac{4x^6+2x^4+4x^2}{2x^6+4x^6-3x^4-4x^2+1}$$

The parameter plane curves for $\zeta = 0$ are plotted in figure 8 and the frequency sensitivities are plotted in figure 9. An M-locus is shown in figure 8 for a saturating amplifier with linear portion gain 70 and α arbitrarily chosen as 10. From the intersection with the $\zeta = 0$ curve, $\beta = 45$ from which $x = 0.174$. The frequency $f = 0.174/2\pi RC$.

Consider the details of a possible design analysis in the parameter plane. An oscillator is to be built using an amplifier with a linear portion gain of 70. A phase shift high pass configuration has been chosen. It is desired to minimize sensitivity to R_3 in this configuration (figure 2). What should α be and what, if any is the effect on the sensitivities with respect to the other parameters if α is changed from an arbitrary reference at $\alpha = 1$?

The sensitivities are calculated from the coefficients of the characteristic equation. It is seen that

$$S_x^{\omega} = -\frac{1}{2} S_x^{\alpha_2}, \text{ since } S_x^{\alpha_o} = 0$$

The sensitivities at $\alpha = 1$ are

$$S_{R_1}^{\omega} = -0.416, \quad S_{R_2}^{\omega} = -0.333, \quad S_{R_3}^{\omega} = -0.250$$

$$S_{C_1}^{\omega} = -0.250, \quad S_{C_2}^{\omega} = -0.333, \quad S_{C_3}^{\omega} = -0.416$$

The maximum possible sensitivity with respect to any single parameter is 0.5. From figure 9, sensitivity with respect to $R_3(\alpha)$ is minimized by making α as small as possible. Reference to figure 8 shows immediately that β increases as α is decreased, hence the linear portion gain places a lower bound on α. A suitable value of α is 0.3 for which $\beta = 55$ and $S_{\alpha}^{x} = -0.115$. The sensitivities at $\alpha = 0.3$ are

$$S_{R_1}^{\omega} = -0.462, \quad S_{R_2}^{\omega} = -0.423, \quad S_{R_3}^{\omega} = -0.115$$

$$S_{C_1}^{\omega} = -0.295, \quad S_{C_2}^{\omega} = -0.333, \quad S_{C_3}^{\omega} = -0.372$$

It is worthy of note that minimizing sensitivity with respect to one parameter is accompanied by the increase in sensitivity with respect to three others, apparently yet another application of the universal principle that you can't get something for nothing!

Reference to figures 8 and 9 also shows immediately, that, in the problem above, second thought should be given to the choice of a high pass phase shift network. Any of the other networks, except the Wien network, would have provided a minimization of sensitivity to terminal resistance much lower than the 0.115 achieved for $\beta = 55$. The phase shift-low pass curve in figure 9 shows that $S_{R_2}^{\omega} < 0.115$ for $\alpha > 2.5$. The value of R_3 is very likely determined by the amplifier specified. If it is suitable to choose $R < R_3/\alpha$ for the required frequency then a change to another configuration is recommended. If this choice is not suitable (C may be very large) then the phase shift high pass configuration may be the only one capable of handling the low input impedance with a low sensitivity.

Many comparisons are possible from figures 8 and 9. In general, it is immediately seen which configurations can meet a sensitivity specification. The network parameters to provide the specified sensitivity are indicated and the required amplifier gain is known. The M-locus for the amplifier indicates whether a configuration is feasible since there may not be an intersection with the $\zeta = 0$ line or the sensitivity specification may demand an impossibly high gain. Such a set of curves can be drawn for any parameter of interest using other fixed values of parameters as required.

AMPLITUDE SENSITIVITY

Since the effective gain of the non=linear amplifier is said to be a monotone decreasing function of "amplitude", and since the effective gain is specified by the intersection of the M-locus with the $\zeta = 0$ curve in the parameter plane, there is a temptation to speak loosely in terms of an amplitude sensitivity. However, the idea of amplitude in non-linear oscillations needs better definition.

If the amplifier gain is regulated such that the end of the M-locus is just inside the unstable region, the harmonic content of the wave form will be small and the amplitude of the fundamental component will be very close to the maximum excursion of the output waveform. If the feedback network is of a low pass configuration then the conditions for application of the describing function may be satisfied and the amplitude of the waveform may be derived from the solution of the transcendental describing function, using the gain specified for the intersection and the linear portion gain specified. However, several of the networks have high pass characteristics and there can be no justification for applying the describing function and deriving an amplitude sensitivity expression therefrom. In general, it follows from the chain rule that

$$S^{\beta}_{\omega} = S^{\beta}_{A} \cdot S^{A}_{\omega}$$

where S^{β}_{ω} is determined from the parametric equations of the ζ = curve and S^{A}_{ω} is the amplitude sensitivity to change in frequency. The derivation of S^{β}_{ω} depends entirely on the dependence of effective gain on amplitude.

CONCLUSIONS

Parameter plane analysis and sensitivity analysis for the frequency characteristics of active networks can be carried out quite readily for oscillators of second and third order configurations without resort to lengthy and generalized parameter plane expressions. Comparison of the $\zeta = 0$ curves and frequency sensitivity curves gives a wide view of alternative structures, and choice of configuration to meet a priori specifications of frequency and frequency sensitivity can be easily made. The technique employed suggests application to sensitivity of frequency characteristics of active and passive filters.

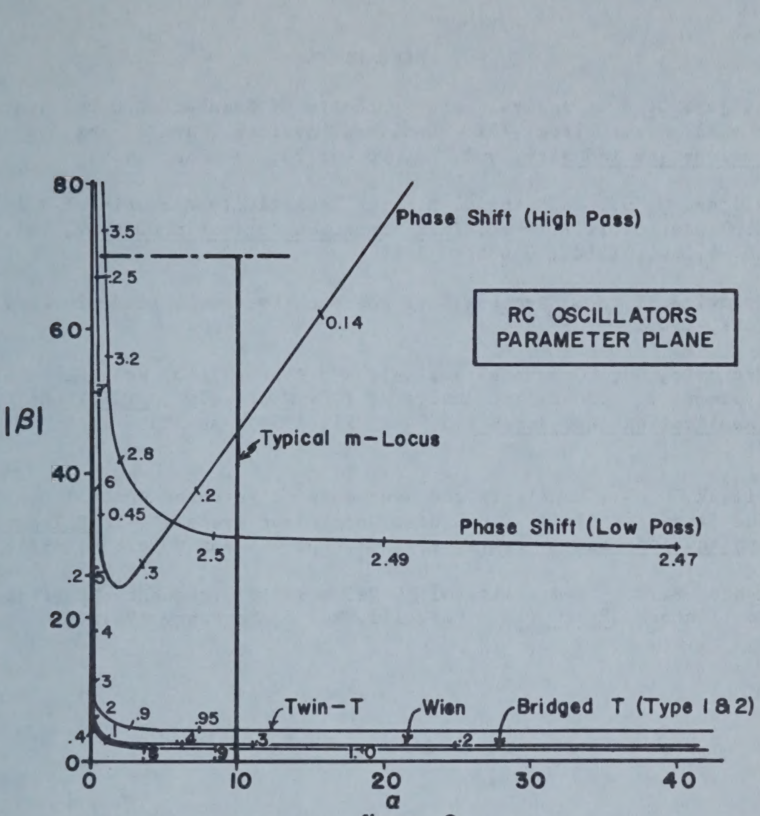


figure 8

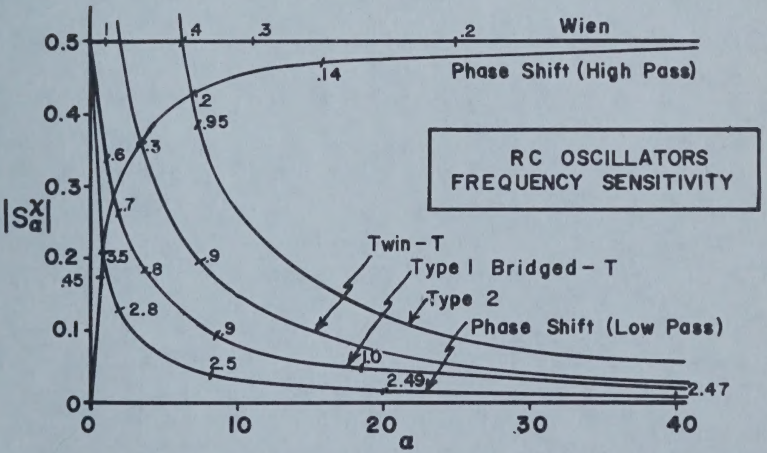


figure 9

REFERENCES

1. Siljak, D. D., "Analysis and Synthesis of Feedback Control Systems in the Parameter Plane, III - Nonlinear Systems", IEEE Trans. on Applications and Industry, vol. 83, pp.466-73, November 1964.

2. Siljak, D. D., and Stojic, M. R., "Sensitivity Analysis of Self Excited Nonlinear Oscillations", IEEE Trans. on Automatic Control, vol. 10, no. 4, pp. 413-20, October, 1965.

3. Horowitz, I. M., "Synthesis of Feedback Systems", Academic Press, New York, 1963.

4. Mitrovic, D., "Graphical Analysis and Synthesis of Feedback Control Systems, I - Theory and Analysis, II - Synthesis", AIEE Trans. pt. II (Applications and Industry), vol. 77, 1958 (Jan. 1959 section), pp 476-96.

5. Siljak, D. D., "Analysis and Synthesis of Feedback Control Systems In The Parameter Plane, I - Linear Continuous Systems", IEEE Trans, on Applications and Industry, vol. 83, pp. 449-58, November, 1964.

6. Mehta, V. B., "Comparison of RC Networks for Frequency Stability in Oscillators, Proc. IEE, Vol. 112, No. 2, February 1965.

WORST CASE GAIN SENSITIVITY WITH ZERO PHASE SENSITIVITY IN ACTIVE RC NETWORK SYNTHESIS

H.E.JONES and B.A.SHENOI
Department Of Electrical Engineering
University Of Minnesota, Minneapolis. Minn.

ABSTRACT

The change in only the magnitude of a network function or the pole vectors due to an infinitely small change in the control parameter K has been usually considered in studying the sensitivity of an active network. In this paper, K is assumed to increase or decrease from its nominal value by finite amounts. A worst case synthesis procedure, based on a new polynomial decomposition, is presented for designing an active RC network, such that the maximum and minimum modulus of the gain function at a prescribed frequency lies within specified limits and also the phase remains unchanged. Example of a transistor amplifier is worked out, including experimental results.

I. STATEMENT OF THE PROBLEM AND ANALYSIS

Though sensitivity has been duly recognised as one of the main criteria when RC-active networks are designed by modern synthesis procedures, various definitions of sensitivity have been used by different authors(1,2,5,6,7) and there seems to be no common understanding about the relative merits and practical use of these mathematical definitions.(4) In all cases, an infinitely small change is assumed in the parameter K with respect to which the gain sensitivity, root sensitivity, coefficient sensitivity, polynomial sensitivity etc. are defined. In practice, however, any device parameter that directly influences K, like the s.c. current gain of a transistor may be much different from the manufacturer's rating; and due to changes in bias currents or temperature over wide range above or below the dc value and ambient temperature respectively, the maximum and minimum value of the parameter to be considered for practical design may bear a very large ratio. A synthesis procedure that provides for such finite increments and decrements from the nominal value of K, within the assumed bounds and assures that the overall gain function at a prescribed frequency- rather than the location of the poles alone- does not exceed a prescribed tolerance, meets the needs of practical design and applications more directly. One such 'worst case' synthesis procedure is presented in this paper. It is so developed that the phase at the prescribed frequency remains the same as K varies over large but finite limits. The concept of phase sensitivity though not new in the existing literature on sensitivity, has hardly been used as a design criteria either alone or in conjunction with magnitude sensitivity. The method described in the following section, which assures a zero phase sensitivity in the active RC network is therefore very novel, and maintaining a zero phase sensitivity as the gain of the amplifier or the transistors varies over wide limits is a great advantage in many applications, e.g. in maintaining the center frequency of an I.F. bandpass network, in which the gain can be independantly adjusted by A.G.C.

Consider the example of realizing a current ratio transfer function (1) having a pair of complex conjugate poles s_a and $\bar{s}_a$ lying in the left half s-plane and no more than two zeros anywhere except the positive real axis. Higher order functions are realizable by cascading such networks as the one shown in Fig.1 that realizes (1).

$$G(s) = \left.\frac{I_2}{I_1}\right|_{V_2=0} = \frac{H\,N(s)}{D(s)} = \frac{H\,N(s)}{s^2 + as + b^2} \tag{1}$$

In terms of the s.c admittance functions of the two RC networks N_a and N_b and the parameter α_f of the current controlled current source, we obtain (2).

$$G(s) = \frac{-\alpha_f\, y_{21b}}{y_{22a} + y_{11b} - \alpha_f\, y_{12a}} \tag{2}$$

The denominator polynomial $D(s) = s^2 + as + b^2$ is to be decomposed into the form (3),

$$D(s) = A_o(s^2 + cs + d^2) + K_o(s^2 + \alpha s + \beta^2) = A_oA(s) + K_oB(s) \tag{3}$$

such that the following conditions are satisfied :

i) the zeros of $A(s) = (s^2 + cs + d^2)$ are simple and negative real

ii) the zeros of $B(s) = (s^2 + \alpha s + \beta^2)$ are complex conjugate and lie in the left half of the s-plane

iii) when the control parameter varies from its nominal value K_o between the limits $K_M = K_o + \Delta K_1$ and $K_m = K_o - \Delta K_2$, defined by $\gamma = (K_M/K_m)$, the magnitude of the corresponding gain functions $|G_{max.}|$ and $|G_{min.}|$ at a prescribed frequency $j\omega_o$ lies within a desired limit specified by

$$\eta = \left|\frac{G_{max.}}{G_{min.}}\right|$$

iv) the $\arg.G(j\omega_o)$ remains constant for any magnitude of deviation in K_o.

Considering a polynomial P(s,K) that satisfies the first two conditions,

$$P(s,K) = A_o(s^2 + cs + d^2) + K(s^2 + \alpha s + \beta^2) \tag{4}$$

it is obvious that as K increases from 0 to ∞, the root locus starts from the two negative real zeros of A(s), breaks off into the complex plane for some value of $K < K_o$, traverses the path of a circle that passes through s_a and $\bar{s}_a$ when $K = K_o$ and finally ends on the zeros of B(s). It is assumed here onwards, that D(s) and N(s) are polynomials obtained when the prescribed frequency ω_o has been normalized to unity. Then

$$\arg.D(j1) = \arg.(b^2 - 1 + ja) \tag{5}$$

and
$$\arg.P(j1) = \arg.\left[A_o(d^2 - 1 + jc) + K(\beta^2 - 1 + j\alpha)\right] \tag{6}$$

If arg.D(j1) = arg.P(j1) for all K, then it is both necessary and sufficient that arg.D(j1) = arg.A(j1) = arg.B(j1) which means

$$\frac{b^2 - 1}{a} = \frac{d^2 - 1}{c} = \frac{\beta^2 - 1}{\alpha} \tag{7}$$

A geometric interpretation of (7) leads to the result that the root locus of (4) after it breaks off into the complex plane is a part of a circle with its center at $-(1 - 4b^2)/2a$ and radius $\sqrt{[(2a + 1 - 4b^2)/2a]}$ and the circle must pass through the normalized frequencies $\pm j\omega_o = \pm j1$ in the s-plane. A few circles when s_a and $s_b = \bar{s}_a$ are specified at different locations are shown in Fig.2. A zero of B(s) can always be chosen on the circle between s_a and j1 along with its conjugate lying on the same circle between $\bar{s}_a$ and -j1. When $K = K_m$ the roots lie on the circle to the left of the zeros s_a and $\bar{s}_a$ and $|G(j1)|$ decreases to $|G(j1)_{min.}|$ and when K increases to K_M the roots move from s_a and $\bar{s}_a$ along the circle to the right towards the zeros of B(s) so that G(j1) increases to $|G(j1)_{max.}|$. Let G_o be identified as the gain function G(s) when K has the nominal value K_o, at which $G_o(s)$ becomes the given function(1).

$$G_o(s) = \frac{K_o N(s)}{A_o A(s) + K_o(s^2 + \alpha s + \beta^2)} \tag{8}$$

$$\left(\frac{G_{max.}}{G_{min.}}\right) = \left(\frac{K_M}{K_m}\right) \frac{A_o A(s) + K_m(s^2 + \alpha s + \beta^2)}{A_o A(s) + K_M(s^2 + \alpha s + \beta^2)} \tag{9}$$

Since $A_o A(s) = (s^2 + as + b^2) - K_o(s^2 + \alpha s + \beta^2)$, the above equation becomes

$$\left(\frac{G_{max.}}{G_{min.}}\right) = \left(\frac{K_M}{K_m}\right) \frac{(s^2 + as + b^2) + (K_m - K_o)(s^2 + \alpha s + \beta^2)}{(s^2 + as + b^2) + (K_M - K_o)(s^2 + \alpha s + \beta^2)} \tag{10}$$

Now the argument condition $\frac{\alpha}{a}(b^2 - 1) = (\beta^2 - 1)$ from (7) is substituted in (10) and its absolute magnitude is evaluated at the normalized frequency j1, giving(11).

$$\left|\frac{G(j1)_{max.}}{G(j1)_{min.}}\right| = \left(\frac{K_M}{K_m}\right) \frac{1 + \frac{\alpha}{a}(K_m - K_o)}{1 + \frac{\alpha}{a}(K_M - K_o)} \tag{11}$$

After some algebraic simplification, we get (12) and (13) shown below:

$$\frac{K_M}{K_m} = \left|\frac{G(j1)_{max.}}{G(j1)_{min.}}\right| \frac{1 - \left|\frac{G(j1)_{min.}}{G_o(j1)}\right| \left(\frac{K_o\alpha}{a}\right)}{1 - \left|\frac{G(j1)_{max}}{G_o(j1)}\right| \left(\frac{K_o\alpha}{a}\right)} = \gamma \tag{12}$$

$$\left(\frac{K_o\alpha}{a}\right) = \frac{\frac{K_M}{K_m}\left|\frac{G(j1)_{min.}}{G(j1)_{max.}}\right| - 1}{\left|\frac{G(j1)_{min.}}{G_o(j1)}\right|\left(\frac{K_M}{K_m} - 1\right)} = \frac{\frac{\gamma}{\eta} - 1}{\left|\frac{G(j1)_{min.}}{G_o(j1)}\right|(\gamma - 1)} \tag{13}$$

Since $\left(\frac{K_o\alpha}{a}\right) > 0$ and $\gamma > 1$, we find $\gamma > \eta$ from (13). Therefore the tolerance on the gain function is limited to a smaller value than the tolerance on the control parameter K. This is an additional attractive feature of this analysis and is clearly substantiated by experimental results.

II SYNTHESIS PROCEDURE

For a given ratio γ and η, $\left(\frac{K_o\alpha}{a}\right)$ is fixed from (13). From the given polynomial $D(s) = (s^2 + as + b^2)$, $(K_o\alpha)$ and also $\frac{b^2 - 1}{a}$ are determined. From (7), the other parameters are selected as described in the illustrative design. Then A_o is calculated from (3) to complete the polynomial decomposition. The RC two-port network N_a is synthesised to realize (14) and (15) in which $Q(s) = (s + \sigma_1)(s + \sigma_2)$ is a quadratic polynomial chosen such that (14) is an RC admittance function. H_a is the multiplier constant actually realized by N_a and is evaluated after the synthesis of N_a is completed.

$$y_{22a} = \frac{A_o A(s)}{Q(s)} \tag{14}$$

and
$$-y_{21a} = \frac{H_a B(s)}{Q(s)} \tag{15}$$

From (2)it is now clear that y_{11b} must be proportinal to y_{22a} by a real constant say h. This restraint is used to advantage for adjusting the value of α_f to a desired value which depends on the transistors chosen for the circuit. Knowing the value of α_f, the value of h is calculated from (16).

$$K_o = \frac{\alpha_f H_a}{1 + h} \tag{16}$$

Identifying the two s.c. admittance functions (17) and (18),

$$y_{11b} = hy_{22a} = \frac{hA_o A(s)}{Q(s)} \tag{17}$$

and
$$-y_{21b} = \frac{H_b N(s)}{Q(s)} \tag{18}$$

the network N_b is synthesized, H_b being the actual value evaluated after the synthesis is completed. The overall gain G(s) of the current amplifier thus realized is given by (19).

$$G(s) = \left(\frac{K_o H_b}{H_a}\right) \frac{N(s)}{D(s)} \tag{19}$$

In (19), K_o is the parameter determined by the polynomial decomposition of the given D(s). A large value of H_b is desirable but this can not be achieved with an a' priori control in the synthesis of N_b. However, it is easier to obtain a low value of H_a, e.g., by choosing Guillemin's parallel ladders method to synthesize N_a and resistive padding whenever possible. The required value of h is also reduced and therefore the s.c. input impedance of N_b is increased compared to the s.c. output impedance of N_b as indicated by (17). Thus a low value of H_a results in a higher gain level in G(s).

III EXPERIMENTAL RESULTS

A bandpass amplifier with a current ratio specified by a normalized

$$G(s = H\frac{s}{s^2 + 0.25\,s + 1}$$

function shown above is chosen for experimental verification of the above theory. Let it be required that the actual center frequency lie at 20,000 radians/sec. (i.e.3.17 kc/s) and remain invariant even when the amplitude of the function varies by ± 20 % due to a temperature change from 25°C to 80°C. In a practical specification like this,where the type of transistors or the controlled sources to be chosen and the expected deviation or allowable tolerance in their parameters have not been specified, we choose arbitrarily a transistor (2N 3567 that has a minimum h_{fe} of 40 at 25°C) for the purpose. From manufacturer's specifications, it is noted that a 3 : 1 spread in its h_{fe} may be expected at 25°C. This is the initial tolerance in its s.c. current gain,though the particular transistor may deviate from the manufacturer's rating by a smaller amount. Allowing for an additional 2 : 1 spread (which is perhaps a pessimistic estimate in practical case) due to the temperature variation from 25°C to 80°C, we assume a value of $\gamma = 6$ for design calculations. Thus we require that $\eta = \frac{1.2}{0.8} = 1.5$ for $\gamma = 6$ and the phase sensitivity to be zero as the overall gain amplitude varies by ± 20 %. With these values,we get$\left(\frac{K_o\alpha}{a}\right) = 0.75$ from(13). In choosing the two negative real zeros of A(s), a large distance between them leads to a large value for the required gain of the controlled source. Arbitrarily an approximate 2 : 1 separation is chosen and A(s) is thus formed as

$$A(s) = A_o(s+\frac{2}{3}d)(s+\frac{3}{2}d)$$

But to satisfy the argument condition of (7) it is necessary that $d^2 = 1$. Hence, the decomposition of the denominator $D(s) = (s^2 + 0.25s + 1)$ reduces to the form

$$D(s) = A_o(s+\frac{2}{3})(s+\frac{3}{2}) + K_o(s^2 + \alpha s + \beta^2)$$

Equating coefficients of these two polynomials, the values of A_o, K_o, α and β^2 are calculated. Thus the decomposition of D(s) is

$$D(s) = (s^2+0.25s+1) = 0.033(s+2/3)(s+3/2) + 0.967(s^2 + 0.194\,s + 1)$$

Choosing Q(s) = (s +1)(s +2) we have

$$y_{22a} = \frac{0.033(s+2/3)(s+3/2)}{(s+1)(s+2)}$$

$$-y_{21a} = \frac{H_a\,(s^2 + 0.194\,s + 1)}{(s+1)(s+2)}$$

and

$$y_{11b} = h\,y_{22a}$$

$$-y_{21b} = \frac{H_b\,s}{(s+1)(s+2)}$$

The admittance levels of networks N_a and N_b can now be adjusted by different factors such that after restoring the frequency scale, the input and output impedances have desirable values,say from considerations of proper bias currents and/or output impedance level for the transistors chosen. The admittance levels also determine the ratio between y_{22a} and y_{11b}. When the ratio(K_o/ H_a) is considered as a fixed constant, (16) indicates it as the minimum value of α_f to be realized by the controlled source. From this value,it is found that one transistor (2N3567) is not sufficient; therefore a Darlington pair is chosen for the controlled source. Also a common base input stage is connected in cascade with it to reduce the input impedance to a sufficiently small value compared to that of N_a and thus approach the condition that the input current to network N_a is zero in Fig.1. The capacitor values are adjusted for restoring the center frequency to 20,000 radians/sec. The gain constant H of the resulting amplifier network which is shown in Fig.3 is calculated as 0.774. This should give a gain of approximately 9 db. at the center frequency.

The network was constructed with standard valued elements(±5% capacitors and ± 1% resistors). The nominal response [curve(a)] measured at 25°C and shown in Fig.4 shows that the center frequency is about 20,8000 radians/sec. and the damping factor is about 0.15 rather than the specified value of 0.125 . The current gain of the controlled source was found to be approximately 800 at 25°C ambient. When this temperature was increased to 80°C, the amplitude response differs from that of the previous case by less than 0.1 db. at the center frequency and almost overlaps at other frequencies[curve(c)]. Next, by shunting across the 36 KΩ collector resistor, the value of the current gain of thecontrolled source was reduced by 6 db., whereby it corresponds to the design minimum K_m. The measurement of the magnitude response at 25°C indicates that the maximum deviation in the response is about 2.7 db.[curve(b)]. The total deviation is therefore about 2.8 db. which is less than the value of 3.5 db. corresponding to ± 20 % specified. The gain of the controlled source was reduced by an additional 14 db. to verify that the zero phase sensitivity is maintained even in this case. It is found that the center frequency does not observably change, although the maximum gain of the amplifier is reduced by about 8 db. [curve(d)]. The test results thus agree reasonably well with the specifications. The principal cause of the slight deviations in the center frequency and damping coefficient is most likely due to the realized location of the complex zero pair of the RC network N_a, which appears to be, however, within expected deviation due to component tolerances. The test results have sufficiently demonstrated the main features of the synthesis procedure, viz.

A. The phase sensitivity to controlled source variations is essentially zero.

B. The magnitude varies by an amount close to the prescribed value over the expected range of variation in the controlled source gain.

IV ACKNOWLEDGEMENT

The support of the above research work received through an NSF Traineeship granted to the first author and an NSF grant (No. GP-3017) granted to the second author is gratefully acknowledged.

V REFERENCES

1. I.M.Horowitz,"Optimization of Negative Impedance Conversion Methods of Active RC Synthesis", IRE Trans. CT-6, pp.296-303. Sept.1959.
2. D.A.Calahan,"Sensitivity Minimization in Active RC Synthesis",IRE Trans. CT-9,pp.38-42.March 1962
3. B.A.Shenoi,"A General Synthesis of Tunnel Diode Networks and Sensitivity Minimization",Proc.NEC, vol.18,pp.114-126. 1962.
4. B.A.Shenoi and S.C.Dutta Roy,"Comparison of Herbst and Horowitz Polynomial Decomposition in RC Active Filter Synthesis", (to appear in Proc.IEE).
5. S.S.Hakim," **Synthesis** of RC Active Filters with Prescribed Pole Sensitivity" Proc.IEE,vol.112,pp.2235-2242,Dec.1965
6. N.M.Herbst,"Optimization of Pole Sensitivity in RC Networks", Cornell University Research Report EE 569,1963
7. J.G.Truxal and I.M.Horowitz,"Sensitivity Considerations in Active Network Synthesis",Proceedings of the 2nd Midwest Symposium on Circuit Theory, Michigan,1956.

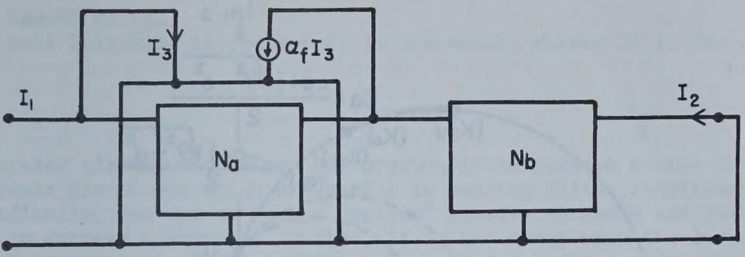

I3
αf I3
I1
I2
Na
Nb

Fig.1 An Active RC Network Embedding a Current Controlled Current Source

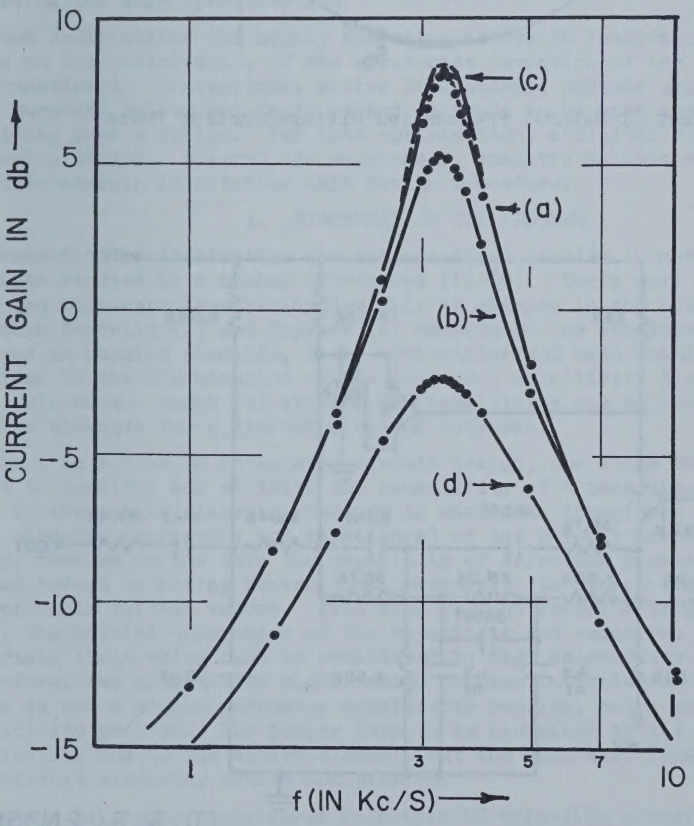


Fig.4 Bandpass Characteristics Realized by the Network of Fig.3

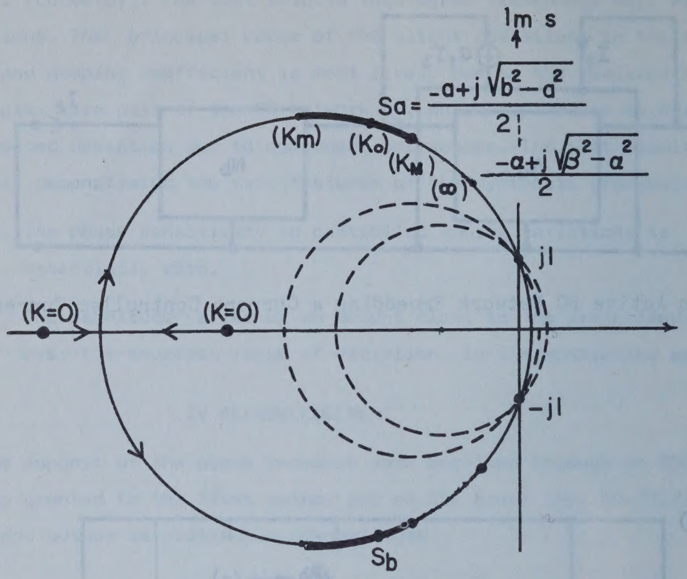


Fig.2 Root Loci Of Natural Frequencies Giving Constant Phase

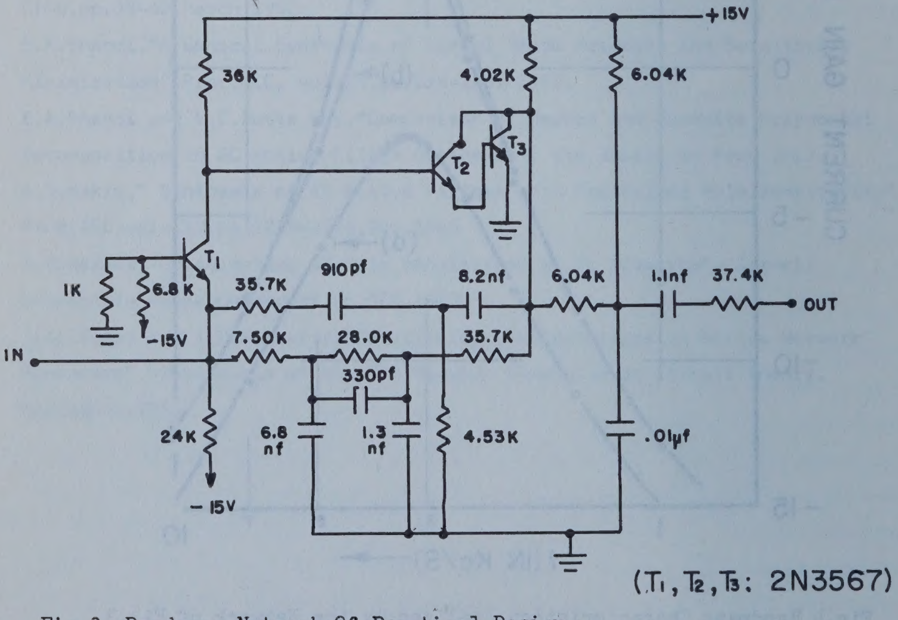

+15V
36K
4.02K
6.04K
T2
T3
T1
1K
6.8K
-15V
35.7K
910pf
8.2nf
6.04K
1.1nf
37.4K
OUT
IN
7.50K
28.0K
35.7K
330pf
24K
6.8 nf
1.3 nf
4.53K
.01μf
-15V
(T1, T2, T3: 2N3567)

Fig.3 Bandpass Network Of Practical Design

SENSITIVITY MINIMIZATION IN ACTIVE RC INTEGRATED CIRCUIT DESIGN

SAMUEL C. LEE
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

ABSTRACT

Integrated circuit technology has profoundly influenced active RC network synthesis procedures which are useful in meeting filter requirements. Specifically, one may no longer "select" passive elements and interconnect them to achieve a network. Rather, it is incumbent upon the designer to accept some or all of the initial values of the passive components which are made simultaneously either in silicon or thin film technology.

Accordingly, although some passive components may be "trimmed to value", the sensitivity problem in active RC integrated circuits is not a problem of minimizing the sensitivity due to the single active element, but rather a multiparameter sensitivity problem. This study emphasizes sensitivity minimization with respect to all the passive and active elements. Active RC realizations using non-series-parallel networks, in general, are found to have smaller multiparameter sensitivity than the conventional realizations achieving the same specification.

Optimum realizations for highly selective active RC integrated circuit design based on the minimization of the worst-case deviation of the (high Q) poles are considered. Conventional active RC synthesis methods are examined for this purpose; Sallen and Key's method is found to be most appropriate for obtaining such a design. For this optimization, a digital computer is generally needed. However, in many cases, analytic solutions may be obtained. A simple example illustrates this design procedure.

I. STATEMENT OF THE PROBLEM

The sensitivity minimization (or optimization) problem in active RC synthesis has been studied by a number of authors [1]-[9]. Their work dealt mainly with the pole-zero sensitivity due only to changes in the active element. Although Horowitz [1] and Sipress [6] considered the sensitivity due to changes in passive elements, their motivations and main results were still confined to the minimization of the pole-zero sensitivity due to active element tolerance. Gaash [9] studied the sensitivity due to both active and passive elements for a limited class of networks.

However, in active RC integrated circuit design, one finds that it is insufficient to consider and minimize the sensitivity of a prescribed network function to the active elements, because in making an integrated circuit, the passive elements (resistors and capacitors) of the circuit are built simultaneously; thus we do not have the capability of selecting passive components as we had before in making conventional networks. In other words, we must often accept their initial values. Even with present resistor trimming technology [10], the initial tolerances of the capacitors and resistors are still within a certain limit which must be considered in high selectivity filter design. Therefore, the sensitivity minimization problem in active RC integrated circuits is not a single parameter sensitivity problem, but a multiparameter sensitivity problem. The proper term to be minimized is not the pole-zero sensitivity due to the active element, but the pole-zero sensitivity due to all circuit elements, active and passive.

The three tolerances considered in active RC thin-film integrated circuits are initial tolerance, temperature tolerance and aging tolerance. With presently available technology including silicon beam-leaded active devices and thin film resistors which track with temperature, the active elements, such as operational amplifiers and negative impedance converters, can be made fairly independent of temperature variation. Using presently available resistor and capacitor temperature coefficient cancellation, (the temperature coefficient of the resistors is made approximately equal but opposite in sign to that of the capacitors) the temperature tolerance of a tantalum thin film circuit RC-product can be made approximately ±25 ppm/°C [10]. Therefore, the frequency response deviation due to temperature tolerance, is nearly equivalent to a frequency scaling factor of $1 \pm (25\ \text{ppm}/°\text{C})\cdot\Delta T$. Even this excellent matching capability can be offset by keeping the circuit in a constant temperature oven; designing (bandpass) filters at lower center frequencies by premodulating the input signal [11]; or using automatic frequency control [12].

In this paper, attention is focussed on the frequency response deviation due to initial tolerance. The aging tolerances of passive and active elements are approximately of the same magnitudes as their initial tolerances. Therefore to minimize one also minimizes the other.

There are two approaches to multiparameter sensitivity and network function deviation problems in active RC integrated circuits:

1. Multiparameter Sensitivity Minimization. This attempts to minimize the multiparameter sensitivity of an integrated circuit, and is based on the assumption that all the circuit component tolerances are small and nearly equal. This problem will be studied in Section II.
2. Worst-Case Deviation Minimization. The quantity to be minimized here is the frequency response deviation which is the summation of the products of the sensitivity functions and the element tolerances, where the element tolerances are assumed to be small but not equal. This is a multiparameter deviation problem. It is important to distinguish between the above problem and this one. The former only concerns sensitivity; the latter also considers element tolerances. We often choose the "worst-case" deviation as the quantity to be minimized for the following reason. In most practical design problems, the specification includes a nominal (or desired) response characteristic with upper and lower bounds. The worst-case response deviation should be within the specified bounds. A study of the optimum realization based on the minimization of worst-case pole deviation is given in Section IV.

To show that the conventional active RC sensitivity minimization criteria (based on the assumption that the active element is the dominant varying parameter) are inapplicable to optimum-sensitivity integrated circuit design, consider an example given by Hakim [8]:

Example 1. Assume

$$K_i(s) = \frac{I_2}{I_1} = \frac{K_i(0)}{s^2 + \sqrt{2}\,s + 1} \tag{1}$$

From the realization (Fig. 1), the complex poles are functions of μ, R_1, R_2, R_3, R_4, C_1 and C_2. The sensitivity of the complex pole $p_c = -0.707 + j0.707$ with respect to these elements are tabulated in Table 1 below.* Assume that the initial tolerances of the capacitors, resistors and gain constant μ are

---

* The sensitivities are calculated from Hakim's network where the element values with only two or three significant figures were used.

$\pm 5\times10^{-3}$, $\pm 2\times10^{-4}$ and $\pm 2\times10^{-4}$ respectively, which may be typical of hybrid thin film integrated circuits. Although all of the resultant deviations are small compared to the real part of p_c in this example, nonetheless it is significant to compare the resultant pole deviations due to individual elements. It is clear that sensitivity to both passive and active elements must be simultaneously considered, as can be seen by comparing the deviations in the fourth column of Table 1.

Table 1

e_i	$S_{e_i}^{p_c}$	Element Tolerance $\pm\varepsilon_i$	$dp_c = S_{e_i}^{p_c}\cdot\pm\varepsilon_i$
μ	$0.030 + j0.0670$	$\pm 2\times10^{-4}$	$\pm(0.60+j1.340)\cdot10^{-5}$
C_1	$0.16 - j0.51$	$\pm 5\times10^{-3}$	$\pm(0.78-j2.5)\cdot10^{-3}$
C_2	$0.51 - j0.15$	$\pm 5\times10^{-3}$	$\pm(0.267-j0.076)\cdot10^{-2}$
R_1	$0.092 - j0.04$	$\pm 2\times10^{-4}$	$\pm(0.18-j0.083)\cdot10^{-4}$
R_2	$0.47 - j0.15$	$\pm 2\times10^{-4}$	$\pm(0.93-j0.29)\cdot10^{-4}$
R_3	$0.10 - j0.45$	$\pm 2\times10^{-4}$	$\pm(0.20-j0.90)\cdot10^{-4}$
R_4	$0.012 - j0.026$	$\pm 2\times10^{-4}$	$\pm(0.24-j0.51)\cdot10^{-5}$

II. MULTIPARAMETER SENSITIVITY MINIMIZATION

In active RC integrated circuits, the network response deviations due to passive elements may be more significant than those due to active elements. Indeed, the "active elements" considered here are generally comprised of several transistors with their associated feedback circuits for stabilization. Transistors in integrated circuit design are often less costly and smaller than either capacitors or resistors and may thus be used freely. Hence the minimization of sensitivities due to passive elements must be considered. Based on the assumption that the tolerances of all circuit components are small and nearly equal, we may consider the following problem: how should we realize the passive RC networks involved in an active RC network synthesis such that the attenuation, phase and network function sensitivity due to all passive elements can be minimized. To answer this, a study of RC structures has evolved to find general network topologies with minimum two-port parameter sensitivity.

Thin film integrated circuit technology has changed some of the criteria by which networks are evaluated. Microminiaturization permits one to place less emphasis on the number of elements in a network, but to require designs which are fairly insensitive to changes in element values. It is found [13]-[14] that non-series-parallel (NSP) networks (which, in general, contain more elements than the equivalent conventional networks), have smaller multiparameter sensitivity than conventional networks realizing the same specifications. As an example, RC ladder networks (series-parallel networks) are compared with respect to multiparameter sensitivity with NSP networks. It has been shown [13] that NSP networks are always less sensitive than the conventional structures. To compare the multiparameter sensitivities of two networks of different topology realizing the same specifications, one needs to find all the sensitivity functions, $S_{e_i}^{T}$, which is quite tedious. However the physical insight that the NSP realizations, in general, are less sensitive than the conventional realizations with respect to multiparameter sensitivity is that the attenuation, phase and network function sensitivity due to each element of the former are more uniformly distributed among elements than for the latter.

Example 2. Suppose it is desired to realize the open-circuited impedance transfer function

$$Z_{21} = \frac{1.44\ s^3(s^3+17.34\ s^2+97.42\ s+182.28)}{8\ s^5 + 1904.4\ s^4 + 1044\ s^3 + 2688\ s^2 + 2680\ s + 1200} \quad (2)$$

by Linvill's negative impedance converter (NIC) method. We may decompose it into the form

$$Z_{21} = \frac{\dfrac{1.2\ s(s^2+13\ s+42)}{s^4 + 18\ s^3 + 97\ s^2 + 180\ s + 100} \cdot \dfrac{1.2\ s(s+4.34)}{s^2 + 6\ s + 5}}{\dfrac{1.2\ s^4 + 31.6\ s^3 + 242.4\ s^2 + 656\ s + 480}{s^4 + 18\ s^3 + 97\ s^2 + 180\ s + 100} - \dfrac{1.2\ s^2 + 9.2\ s + 12}{s^2 + 6\ s + 5}} \quad (3)$$

$$= \frac{z_{12a} \cdot z_{12b}}{z_{22a} - z_{11b}} \quad (4)$$

which can be realized by the network shown in Fig. 2(a). It is found (see Ref. 13, pp. 44-46 and pp. 123-125) that it can also be realized by a NSP realization as shown in Fig. 2(b).

Since the two realizations all realize (3), the sensitivities due to the active elements of the two networks are the same. It has been shown (see Ref. 13, Chapters V and VI) that the attenuation multiparameter sensitivities of z_{12a} and z_{12b} of the two NSP networks (N_a and N_b of Fig. 2(b)) are less than those of the two corresponding ladder networks (N_a and N_b of Fig. 2(a)), respectively. Following the same lines as the proofs for z_{12a} and z_{12b}, one can show that the multiparameter sensitivities of z_{22a} and z_{11b} of the former are also less than the latter. Hence the attenuation multiparameter sensitivities of Z_{21} of the second (NSP) realization should be less than those of the first, conventional, realization with respect to all passive elements. The phase and network function multiparameter sensitivities of the NSP realization are also found to be less than those of the conventional realization [Ref. 13, pp. 117-118 and 154-155]. As to the minimization of the sensitivity due to the NIC conversion ratio k, we can apply Horowitz's optimum decomposition [1].

The problem next considered is the multiparameter sensitivity minimization in designing high selectivity active RC filters.

III. GENERAL CONSIDERATIONS OF HIGH SELECTIVITY FILTER DESIGN

In most practical high selectivity filter design problems, the design specification contains two parts: the filter nominal characteristics and the sensitivity (or deviation) requirements. They are both very stringent. To meet the former requires only high-accuracy computation of element values and high-precision circuit components. To meet the latter, one must design a network, for which the network response deviation due to all the element tolerances is minimum. This broad problem involves an optimum approximation, an optimum decomposition of the network function, and a network synthesis and deviation minimization problem (see Fig. 3). A brief discussion of the first two is given below. The last one will be considered in the next section.

The first important aspect of any network synthesis is the approximation of the prescribed network characteristic by a realizable mathematical expression. The network function obtained from an approximation of a high performance filter specification of ten contains a large number of poles and zeros including high Q[*] complex poles. Such a network function (assume an open-circuit voltage transfer function) can be realized by one or more passive RC networks and a cascade of active networks, each of which realizes a biquadratic function with a pair of complex poles. This is a mathematical decomposition of a network function. Consider the biquadratic open-circuited voltage transfer function

$$G(s) = H\,\frac{(s-z_1)(s-z_2)}{(s-p_c)\left(s-p_c^*\right)} = \frac{H(s^2+a_1s+a_2)}{s^2+2\sigma_c s+\omega_n^2} = \frac{N(s)}{D(s)} \qquad (5)$$

where H is a constant, $p_c = -\sigma_c + j\omega_c$ and p_c^* is its complex conjugate. $\omega_n^2 = \sigma_c^2 + \omega_c^2$ and it is assumed that $\omega_c \gg \sigma_c$. z_1 and z_2 are two zeros which may have any locations in the s-plane. Suppose that G(s) of (5) is realized by an active RC network having n elements, $e_1, \ldots, e_n$. Then G(s) may be expressed as

$$G(s) = G(e_1, \ldots, e_n, s) = \frac{N(e_1, \ldots, e_n, s)}{D(e_1, \ldots, e_n, s)} \qquad (6)$$

The general form of $S_{e_i}^G$ may be expressed as

$$S_{e_i}^G = H + \frac{\gamma_3 s^3 + \gamma_2 s^2 + \gamma_1 s + \gamma_0}{\left(s^2+2\sigma_c s+\omega_n^2\right)(s+a_1s+a_2)} = H + \frac{F(s)}{\left(s^2+2\sigma_c s+\omega_n^2\right)(s+a_1s+a_2)} \qquad (7)$$

It is known [15] that the correlation between the sensitivity function and the pole-zero sensitivity, $S_{e_i}^p \triangleq \dfrac{dp}{d(\ln e_i)}$; is that

$$S_{e_i}^p = \text{Residue of } S_{e_i}^G \text{ at "} s = p \text{"} \qquad (8)$$

where p may either be a pole or a zero of G(s). Moreover, from the definition of $S_{e_i}^G$, one can show that

$$S_{e_i}^G = \frac{S_{e_i}^{p_c}}{s - p_c} + \frac{S_{e_i}^{p_c^*}}{s - p_c^*} - \frac{S_{e_i}^{z_1}}{s - z_1} - \frac{S_{e_i}^{z_2}}{s - z_2} + S_{e_i}^H \qquad (9)$$

If we assume that changes in G on the imaginary axis at ω_c (in the neighborhood of the pole p_c) are of primary interest, then all terms in (9) except the one involving p_c may be neglected. $S^H_{e_i}$ can be disregarded because changes in H affect only the gain level and, hence, are usually of less consequence in a high selectivity filter. Thus

---

[*] The Q of a complex pole $p_c = -\sigma_c + j\omega_c$ is defined as $Q = \omega_c/2\sigma_c$.

$$S^G_{e_i} \approx \frac{S^{p_c}_{e_i}}{s - p_c} \tag{10}$$

From (7), (8) and (10),

$$S^G_{e_i} \approx \frac{\dfrac{F(p_c)}{j2\omega_c(p_c-z_1)(p_c-z_2)}}{s + \sigma_c - j\omega_c} \tag{11}$$

At $\omega = \omega_c$,

$$S^G_{e_i} \approx Q \cdot \frac{1}{\omega_c^2} \cdot \frac{F(p_c)}{(p_c-z_1)(p_c-z_2)} \tag{12}$$

A pair of complex poles for which $\omega_c \gg \sigma_c$ corresponds to a "peak" in the frequency response at ω_c, in high selectivity filter design. Thus ω_c is essentially predetermined by the specification. Equation (12) tells us that the sensitivity function is proportional to Q. Therefore, in determining an approximation from a given frequency response specification, one should find an approximation such that the network function obtained contains the lowest Q poles consistent with the specification.

Another consequence may be drawn from (12), i.e., the magnitude of (12), which represents changes in network function due to changes in e_i, is inversely proportional to the distances between the complex pole p_c and the zeros. Thus when the network function is decomposed into a product of biquadratic functions, the high Q complex poles should be associated with the zeros in such a way that the distances from the upper half plane complex poles to the respective zeros are made as far self-distant for each subnetwork function as possible.

IV. THE WORST-CASE DEVIATION MINIMIZATION

Having discussed the approximation and optimum decomposition problem based on the multiparameter sensitivity considerations, now consider the network synthesis and deviation minimization of the biquadratic open-circuit voltage transfer function of (5). Suppose G(s) can be realized by an active RC two-port network N containing n elements $e_1,\dots,e_n$, with n_r resistors, n_c capacitors and one active element. The open-circuited voltage transfer function in terms of the circuit components of N is

$$G(s) = f_0 \frac{s^2 + f_1 s + f_2}{s^2 + f_3 s + f_4} \tag{13}$$

where $f_0,\dots,f_4$ are functions of $e_1,\dots,e_n$. Equating (5) and (13), one has

$$f_0(e_1,\dots,e_n) = H \quad \dots \quad f_4(e_1,\dots,e_n) = \omega_n^2 \tag{14}$$

These conditions will be referred to as function coefficient conditions. In solving (14), the values of $e_1, \ldots, e_n$ must be within the fabrication limitations on the elements; for example, for tantalum thin-film resistors and capacitors, the ranges are approximately $10\Omega \leq R \leq 10^6\Omega$ and $100 \text{ pf} \leq C \leq 0.05$ µf. The conditions imposed on the element values limited by the fabrication will be referred to as fabrication limitation conditions. Suppose p is a pole or a zero of G(s) and N is an active RC thin film integrated circuit. Assume $\pm\varepsilon_{r_i}$, $\pm\varepsilon_{c_i}$ and $\pm\varepsilon_k$ are the initial tolerances on the R_i's and C_i's and k, where k is either the gain constant of an operational amplifier or the conversion ratio of a negative impedance converter, respectively. The worst-case deviation of p, WCD(p), caused by the initial tolerances is

$$\text{WCD}(p) \triangleq \sum_{i=1}^{n_r} \left[\left\| S^p_{R_i} \right\| \cdot \left| \pm\varepsilon_{r_i} \right| \right] + \sum_{i=1}^{n_c} \left[\left\| S^p_{C_i} \right\| \cdot \left| \pm\varepsilon_{c_i} \right| \right] + \left\| S^p_k \right\| \cdot \left| \pm\varepsilon_k \right| \quad (15)$$

For a biquadratic voltage transfer function containing a high-Q pole, it has been shown (see Equation (10)), that the sensitivity function is nearly directly proportional to pole-zero sensitivity of p_c. Moreover, the attenuation and phase sensitivities, which are the real and imaginary parts of the network function, respectively, are proportional to it. Thus, the proper term to be minimized in high selectivity filter design is the worst-case deviation of p_c due to all circuit components, namely, WCD(p_c). Now we define:

Definition: Optimum Worst-Case Deviation Realization. An active RC circuit which realizes G(s) of (5) is called an optimum worst-case deviation realization if the element values of the realization are such that the worst-case deviation of p_c is minimized, in addition to meeting function coefficient conditions and fabrication limitation conditions.

Now attention is called to the problem of obtaining optimum tolerance realizations. First, let us examine the active RC synthesis methods which have appeared in the literature. It is found that they can be classified into two categories: in one a network function F(s) (driving-point or transfer immittance function) is synthesized by introducing a polynomial dividing both the numerator and denominator of F(s); in the other F(s) is synthesized directly. For convenience, the former will be referred to as the indirect synthesis and the latter, the direct synthesis. The realizations obtained from the indirect synthesis usually contain more elements and require more function coefficient conditions then those obtained from the direct synthesis that realize the same network function. Furthermore, the function coefficient conditions of the former are complicated which make the assignment on the element values very difficult. Thus, from the sensitivity optimization point of view, it is preferred to use the direct synthesis because it provides network realizations having flexibility (or freedom) of assignment of the element values so that the worst case deviation of p_c can be sufficiently minimized.

Based on the factors:

(1) Accuracy and element economy.

(2) The network realization flexibility with respect to assignment of the element values.

(3) The gain of the active element being finite, among the known open-circuited voltage transfer function active RC synthesis methods [3] [6]

[8] [16]-[19], it is found that Sallen and Key's [19] method* meets these requirements, and is suitable for obtaining optimum deviation realizations. The optimum deviation realization of (5) using configurations in Sallen and Key's table is next considered.

It can be verified that all the sensitivity functions of G with respect to each element** of any circuit given in Sallen and Key's table may be expressed*** as

Case A. N(s) of (6) is 1, s, or s^2.

$$S^{G}_{e_i} = \frac{\alpha_i s + \beta_i}{D(s)} + (1,\ 0,\ \text{or } -1) \tag{16}$$

Case B. N(s) of (6) is $s + 1/\tau$, $s(s+1/\tau)$, or $s^2 + bs + 1$. The sensitivity functions will either be of the form of (16) or,

$$S^{G}_{e_i} = \frac{\alpha_i s + \beta_i}{D(s)} - \frac{\gamma_i s + \delta_i}{N(s)} \tag{17}$$

where D(s) and N(s) are as defined in (6). α_i and γ_i are functions of R_i's, C_i's and K. β_i and δ_i are either 1, 0, K, or K + 1. For example, sensitivity functions of G(s) with respect to each circuit element in terms of α, β, γ, and δ for circuit No. 1 in Sallen and Key's table are given in Table 2. From (16) and (17), using (8), one obtains

$$S^{p_c}_{e_i} = \frac{\alpha_i}{2} + j\left[\frac{\alpha_i}{4Q} - \frac{\beta_i}{2\omega_c}\right] \tag{18}$$

or

$$S^{p_c}_{e_i} = \left[\frac{\alpha_i}{2} - \frac{\omega_c}{N(p_c)}\cdot\gamma_i\right] + j\left[\frac{\alpha_i}{4Q} - \frac{\beta_i}{2\omega_c} - \frac{1}{N(p_c)}(\sigma_c\gamma_i - \delta_i)\right] \tag{19}$$

If G(s) of (5) is normalized by ω_n, as did in Sallen and Key's paper, and $\sigma_c \ll \omega_c$, then (18) and (19) may be reduced to

$$S^{p_c}_{e_i} = \frac{\alpha_i}{2} - j\frac{\beta_i}{2} \tag{18'}$$

$$S^{p_c}_{e_i} = \left[\frac{\alpha_i}{2} - \frac{\gamma_i}{N(p_c)}\right] - j\left[\frac{\beta_i}{2} - \frac{\delta_i}{N(p_c)}\right] \tag{19'}$$

---

* The Q of a complex pole $p_c = -\sigma_c + j\omega_c$ is defined as $Q = \omega_c/2\sigma_c$.

** The variation of K_1 only effects the gain level of G(s); we thus can omit it in the discussion of pole-zero sensitivity.

*** For circuits No. 4, 5, and 12 in Sallen and Key's table, it is assumed that $K + 1 \approx K$ under the condition $\sigma_c \ll \omega_c$. This is an appropriate assumption because from their design formula table, it is found that K_{min} for these three circuits are $\frac{4(1+x) - d^2}{d^2}$. K will be much greater than unity ($K \gg 1$) if $\sigma_c \ll \omega_c$, or the Q of the complex pole is high.

The procedure of optimizing the worst-case deviation with respect to all circuit elements is: find $S^G_{e_i}$ to obtain $S^p_{e_i}$ for all elements and substitute them into (15). Then minimize $WCD(p_c)$ of (15) under the side constraints on the element values, which are the fabrication limitation conditions and the function coefficient conditions given by Sallen and Key. In general, Equation (15) is too complicated to solve analytically. A computer assistance is needed for finding this optimum solution. However, in many cases analytic solutions may be obtained.

Example 3. Suppose we convert the network of Fig. 1 by applying the current source-to-voltage source transformation and using the resistor R_7 as load resistor. Then the open-circuited voltage transfer function of the resultant network is

$$G(s) = \frac{V_2}{V_1} = \frac{R_7 I_2}{R_0 I_1} = \frac{R_7}{R_0} \cdot K_i(s) = \frac{\frac{R_7}{R_0} K_i(0)}{s^2 + \sqrt{2}\, s + 1} \tag{20}$$

Suppose the first circuit in Sallen and Key's catalogue is used to realize (20). This circuit belongs to the general class of "active feedback" circuits. Using Table 2 and applying (16), one can obtain the pole-zero sensitivity of p_c ($p_c = -0.707 + j0.707$) due to each element. Again, assume that the initial tolerances of the capacitors, resistors, and the gain K of the operational amplifier are $\pm 5\times10^{-3}$, $\pm 2\times10^{-4}$ and $\pm 2\times10^{-4}$, respectively. Substitution of the pole-zero sensitivities of p_c into (15) yields

$$\begin{aligned} WCD(p_c) = \frac{1}{2}\Bigg\{ & 5\times10^{-3} \cdot \bigg\{ \Big[\big((R_1+R_2)C_2\big)^2 + \big((R_1+R_2)C_2 - \sqrt{2}\big)^2 \Big]^{\frac{1}{2}} \\ & \Big[\big((1-K)R_1C_1\big)^2 + \big((1-K)R_1C_1 - \sqrt{2}\big)^2 \Big]^{\frac{1}{2}} \bigg\} \\ & + 2\times10^{-4} \cdot \bigg\{ \Big[(R_2C_2)^2 + (R_2C_2 - \sqrt{2})^2 \Big]^{\frac{1}{2}} + \Big[\big(R_1C_2 + (1-K)R_1C_1\big)^2 \\ & + \big(R_1C_2 + (1-K)R_1C_1 - \sqrt{2}\big)^2 \Big]^{\frac{1}{2}} + \Big[(KR_1C_1)^2 + (KR_1C_1)^2 \Big]^{\frac{1}{2}} \bigg\} \Bigg\} \end{aligned} \tag{21}$$

The factors 5×10^{-3} and 2×10^{-4} may be visualized as weighting factors to the functions with which they are associated. From (21), one observes that $WCD(p_c)$ will be sufficiently minimized if $(1-K) \approx K$ (i.e., $K \approx 0.5$) and the value of the product R_2C_2 is designed to be as close as possible to that of R_1C_1 in addition to the design formulas for dissipation factor d given in Sallen and Key's paper [19]. Among the four sets of design parameters associated with the circuit, it is found that the second, namely, using $\rho(=R_1/R_2)$, $T_2(=R_2C_2)$, and the formula for d in group I of [19] is most convenient for achieving such a design. The procedure of designing circuit element values which give an optimum deviation realization is as follows:

1. The worst-case deviation optimization requires that T_2 be close to T_1 and $K \approx 0.5$; following the notation of [19] we let

$$T_2 = T_{K_{min}} = \frac{d}{2(1+\rho)} \tag{22}$$

$$K = K_{min} = 1 - \frac{d^2}{4(1+\rho)} \tag{23}$$

2. Since $d = \sqrt{2}$, from (23), K must be greater than 0.5. From a close observation of (22), (23) and the condition $T_1 \cdot T_2 = 1$ [19], it is found that if K is set close to 0.5, the value of T_2 will sufficiently be close to T_1. If we let $K = 0.51$, $\rho = 0.0204$. Substitution into (22) gives $T_2 = 0.683$. From the condition $T_1 \cdot T_2 = 1$, it follows that $T_1 = 1.464$.
3. If R_1 is set to be 10, then $R_2 = 490.2$, $C_1 = 0.1464$ and $C_2 = 0.001393$. The realization is completed and the sensitivities of p_c due to the elements are calculated and given in Table 3.

Table 3

e_i	$S_{e_i}^{p_c}$	e_i	$S_{e_i}^{p_c}$
C_1	0.3535 - j0.3535	R_2	0.3655 - j0.3415
C_2	0.3535 - j0.3535	K	0.3733 + j0.3733
R_1	0.3415 - j0.3655		

It can be shown that $S_{C_1}^{p_c} + S_{C_2}^{p_c} = -p_c$ and $S_{R_1}^{p_c} + S_{R_2}^{p_c} = -p_c$. Since $S_{C_1}^{p_c} = S_{C_2}^{p_c}$ and their real and imaginary parts are equal, the sum of the $\left|S_{C_1}^{p_c}\right|$ and $\left|S_{C_2}^{p_c}\right|$ of Table 3 is minimum. Similarly, the sum of $\left|S_{R_1}^{p_c}\right|$ and $\left|S_{R_2}^{p_c}\right|$ is nearly minimum. Note that the magnitudes of $S_{C_i}^{p_c}$ are smaller than those of $S_{R_i}^{p_c}$ and $S_K^{p_c}$ as it is desired, since this is an optimum worst-case deviation realization. The WCD(p_c) of the realization is calculated to be 5.195×10^{-3} rads/sec. The WCD(p_c) of the realization of Example 1 is 5.56×10^{-3} rads/sec.

V. CONCLUSION

It has been pointed out that the sensitivity problem in active RC integrated circuit realizations is a multiparameter sensitivity problem where all circuit elements should be considered as parameters. To minimize the multiparameter sensitivity with respect to passive circuit components in active RC synthesis, it is found that the active RC realizations using non-series-parallel RC networks have smaller multiparameter sensitivity than the conventional realizations.

Thin-film circuit design with sensitivity as a design parameter and the optimum worst case deviation realization with respect to all the circuit element initial tolerances have been studied. Conventional active RC synthesis methods are examined. It is found that Sallen and Key's method is most suitable for obtaining such a realization. A simple example has been given to illustrate the method.

ACKNOWLEDGMENT

The author wishes to thank Dr. Ralph W. Wyndrum, Jr. of Bell Telephone Laboratories, Inc., for his invaluable discussions and suggestions.

REFERENCES

1. Horowitz, I. M., "Optimization of Negative-Impedance Conversion Methods of Active RC Synthesis," Trans. IRE CT-6, pp. 293-303, September, 1959.
2. Horowitz, I. M., "Exact Design of Transistor RC Band-Pass Filter with Prescribed Active Parameter Insensitivity," Trans. IRE CT-7, pp. 313-320, September, 1960.
3. Horowitz, I. M., "Optimization Design of Single Stage Gyrator - RC Filters with Prescribed Sensitivity," Trans. IRE CT-8, pp. 88-94, June, 1961.
4. Calahan, D. A., "Notes on the Horowitz Optimization Procedure," IRE Trans. on Circuit Theorem, Vol. CT-7, pp. 352-354, September, 1960.
5. Calahan, D. A., "Sensitivity Minimization in Active RC Synthesis," IRE Trans. on Circuit Theory, Vol. CT-9, No. 1, pp. 38-42, March, 1961.
6. Sipress, J. M., "Synthesis of Active RC Networks," Polytechnic Institute of Brooklyn Electrical Engineering Department, Dissertation, June, 1960.
7. Holt, A. G. J., and Stephenson, F. W., "A Specialization of the Yanagisawa Synthesis Procedure to Obtain RC Active Networks Having Optimum Pole Sensitivity," Proc. IEE, Vol. 112, February, 1965.
8. Hakim, S. S., "Synthesis of RC Active Filters with Prescribed Pole Sensitivity," Proc. IEE, Vol. 112, No. 12, December, 1965.
9. Gaash, A., Pepper, R., and Perderson, D., "Design of Integrable Desensitized Frequency Selective Amplifiers," Digest of Solid-State Circuit Conference, pp. 34-35, February, 1966.
10. Orr, W. H., "Precision Tuning of Thin Film Notch Filter," Digest of the International Solid State Circuits Conference, Philadelphia, pp. 56-57, 1965.
11. Blecher, F., and Hallenbeck, F., "The Transistorized A5 Channel Bank for Broadband Systems," BSTJ, Vol. 41, pp. 321-359, January, 1962.
12. Terman, F. E., "Electronic and Radio Engineering," McGraw Hill Book Co., New York, 4th Edition, 1955.
13. Lee. S. C., "Network Synthesis to Minimize Multiparameter Sensitivity," Ph.D. Thesis, University of Illinois, August, 1965.
14. Lee, S. C., "Non-Series-Parallel Realization of Bisymmetrical Two-Element-Kind Two-Ports," Proceedings of Fourth Annual Allerton Conference on Circuit and System Theory, 1966.
15. Mikulski, J. J., "A Correlation Between Classical and Pole-Zero Sensitivity," Tech. Note No. 5, Contract No. AF49(638)-630, Electrical Engineering Research Laboratory, University of Illinois, 1959.
16. Yanagisawa, T., "RC Active Networks Using Current Type Negative Impedance Converters," IRE Trans. on Circuit Theory, Vol. CT-4, pp. 140-144, September, 1957.
17. Kinariwala, B. K., "Synthesis of Active RC Networks," BSTJ, Vol. 38, p. 1269, September, 1959.
18. Holt. A. G. J., and Sewell, J. I., "Active RC Filters Employing a Single Operational Amplifier to Obtain Biquadratic Responses," Proc. IEE, Vol. 112, No. 12, December, 1965.
19. Sallen, R. P., and Key, E. L., "A Practical Method of Designing RC Active Filters," IRE Trans. on Circuit Theory, Vol. CT-2, pp. 74-85, March, 1955.
20. Balabanian, N., and Patel, B., "Realization of Complex Zeros," IEE Circuit Theory Transactions, Vol. CT-10, No. 2, pp. 299-300, June, 1963.

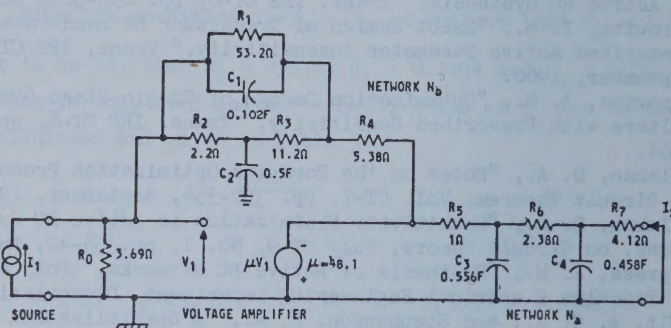
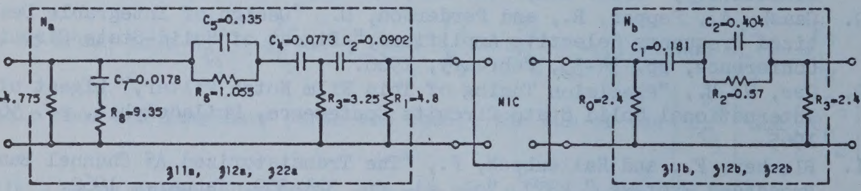
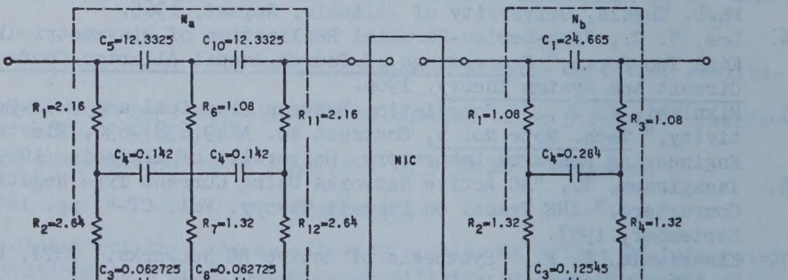

R1
53.2Ω
C1
0.102F
NETWORK Nb
R2
2.2Ω
R3
11.2Ω
R4
5.38Ω
C2
0.5F
R5
1Ω
R6
2.38Ω
R7
4.12Ω
I2
I1
R0
3.69Ω
V1
μV1
μ=48.1
C3
0.556F
C4
0.458F
SOURCE
VOLTAGE AMPLIFIER
NETWORK Na

FIG. 1 HAKIM'S EXAMPLE


Na
C5=0.135
C4=0.0773
C2=0.0902
C7=0.0178
R6=1.055
R9=4.775
R8=9.35
R3=3.25
R1=4.8
NIC
Nb
C2=0.404
C1=0.181
R0=2.4
R2=0.57
R3=2.4
y11a, y12a, y22a
y11b, y12b, y22b

(a) A CONVENTIONAL REALIZATION


Na
C5=12.3325
C10=12.3325
R1=2.16
R6=1.08
R11=2.16
C4=0.142
C4=0.142
NIC
R2=2.64
R7=1.32
R12=2.64
C3=0.062725
C8=0.062725
y11a, y12a, y22a
Nb
C1=24.665
R1=1.08
R3=1.08
C4=0.284
R2=1.32
R4=1.32
C3=0.12545
y11b, y12b, y22b

(b) A NON-SERIES-PARALLEL REALIZATION

FIG. 2 REALIZATIONS OF EXAMPLE 2.

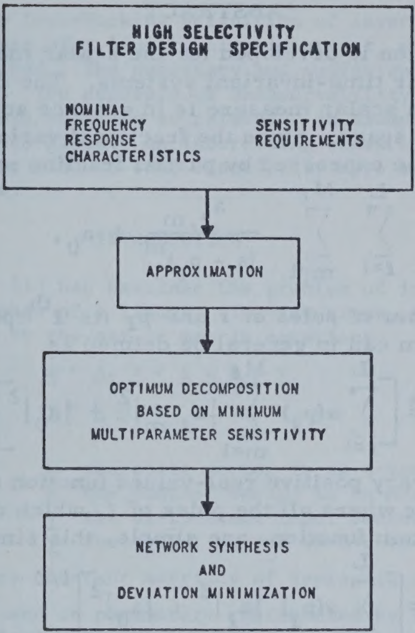
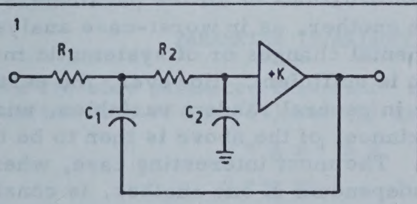

HIGH SELECTIVITY FILTER DESIGN SPECIFICATION
NOMINAL FREQUENCY RESPONSE CHARACTERISTICS
SENSITIVITY REQUIREMENTS
APPROXIMATION
OPTIMUM DECOMPOSITION BASED ON MINIMUM MULTIPARAMETER SENSITIVITY
NETWORK SYNTHESIS AND DEVIATION MINIMIZATION

FIG. 3 A PROCEDURE OF OBTAINING MINIMUM MULTIPARAMETER SENSITIVITY ACTIVE RC FILTER DESIGN FOR HIGH-PERFORMANCE SPECIFICATION.

$$S_{e_i}^{G} = \frac{\alpha_i s + \beta_i}{D(s)} + C_i$$

CONFIGURATION


1
R1
R2
C1
C2
+K

e	α	β	C
R_1	R_2C_2	1	-1
C_1	$(R_1+R_2)\ C_2$	1	-1
R_2	$R_1C_2+(1-K)R_1C_1$	1	-1
C_2	$(1-K)R_1C_1$	1	-1
K	KR_1C_1	0	1

TABLE 2

A NEW DEFINITION OF MULTIPARAMETER SENSITIVITY[*]

R. N. BISWAS
Department of Electrical Engineering,
University of California, Berkeley

ABSTRACT

A new definition is developed for the scalar multiparameter sensitivity of lumped linear time-invariant systems. The definition stems from the fact that any such scalar measure is in essence some kind of norm on the relevant function space. Since the fractional variation of the system function can always be expressed by partial fraction expansion as

$$f(s) = \sum_{\ell=1}^{L} \sum_{m=1}^{M_\ell} \frac{a_{\ell,m}}{(s - p_\ell)^m} + a_0 ,$$

where L is the number of poles of f and p_ℓ its ℓ^{th} pole, having multiplicity M_ℓ, the norm can in general be defined as

$$\|f\| \triangleq \left[\sum_{\ell=1}^{L} \alpha(p_\ell) \sum_{m=1}^{M_\ell} |a_{\ell,m}|^2 + |a_0|^2 \right]^{1/2} ,$$

where α is an arbitrary positive real-valued function of the complex variable. In the case where all the poles of f, which consist of the poles and zeros of the system function, are simple, this simplifies to

$$\|f\| = \left[\sum_{\ell=1}^{L} \alpha(p_\ell) \; |a_\ell|^2 + |a_0|^2 \right]^{1/2}$$

It is then shown that a_ℓ is just the variation in the ℓ^{th} root, and thus the multiparameter sensitivity is directly related to the familiar root sensitivities. In the case of multiple roots, this offers a logical extension of the same notion, and in that sense provides better physical interpretations than most of the existing definitions. Flexibility of the choice of the root weighing function α permits one to tailor the definition to provide a sensitivity index most appropriate to any set of design objectives.

In the case where the perturbations of all the parameters are deterministically related to one another, as in worst-case analysis, or in considering effects of environmental changes or of systematic manufacturing errors, the above definition is sufficient. However, the perturbations of the different parameters are in general random variables, and a stochastic measure, e. g., the variance, of the above is then to be taken as the multiparameter sensitivity. The most interesting case, where the parameter perturbations are independent of one another, is considered in some detail. A simple example illustrates the applicability of this definition to provide a criterion for comparing the performance of different systems realizing the same input-output relation.

---

[*] The research reported herein was supported wholly by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Grant AF-AFOSR-139-66.

EIGENFUNCTION SENSITIVITY AND INVARIANCE

D. C. REDDY
Department of Electrical Engineering
University of Iowa
Iowa City, Iowa

ABSTRACT

Starting with the Cruz-Perkins definition of invariance, the problem of an invariant response of a linear, time-invariant, multivariable, differential system is examined. The necessary and sufficient conditions for such a response are stated. The concepts of unobservability, invariance, eigenfunction sensitivity are shown to be related. An example of a system with an invariant response is given. An interesting class of matrices - circulants - which give rise to some identically zero sensitivity functions are finally introduced.

INTRODUCTION

An earlier paper [1] has examined the problem of invariance and sensitivity of system response of a linear, time-invariant, multivariable differential system defined by the set of matrix equations

$$\dot{\underline{x}} = \underline{A}_1 \underline{x} + \underline{B}\,\underline{u} + \underline{N}\,\underline{v} \qquad (1)$$

and

$$\underline{y} = \underline{c}\,\underline{x}$$

where (2)

$\underline{x}$ the state vector is an n-vector
$\underline{u}$ the control vector is an m-vector
$\underline{v}$ the disturbance input vector is a k-vector
$\underline{y}$ the output vector is a q-vector

and $\underline{A}_1$, $\underline{B}$, $\underline{N}$, and $\underline{c}$ are constant matrices of appropriate dimension. Also these matrices may depend on parameters represented by the r-vector $\underline{\alpha}$, which may vary from its nominal value $\underline{\alpha}_o$.

The object of this paper is to examine the same problem of invariance of system response from yet another point of view. Specifically, to gain an additional insight into this rather difficult problem, extensive use has been made of the notion of eigenfunction sensitivity. This approach reveals certain interesting relationships between the concepts of invariance, unobservability, eigenfunction sensitivity and time-domain sensitivity functions. Also an interesting class of matrices - circulants - which have certain basic properties of invariance are finally introduced.

PARAMETER INVARIANCE

The term "totally parameter invariant" is first defined. Consider the slightly modified form of Eqs. (1) and (2) given below, namely

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{N}\,\underline{v} \qquad (3)$$

$$\underline{y} = \underline{c}\,\underline{x} \qquad (4)$$

and

$$\underline{u} = \underline{F}\,\underline{y} \qquad (5)$$

where

$$\underline{A} = \underline{A}_1 + \underline{B}\,\underline{F}\,\underline{c} \qquad (6)$$

Let the matrix $\underline{A}$ of Eq. (3) become $\underline{A} + \underline{E}$ due to changes in the nominal value of the parameters of the system. The zero input response of the system (i.e., $\underline{v} = \underline{0}$) of Eq. (3) of above is said to be totally parameter invariant with respect to a change of $\underline{A}$ to $\underline{A} + \underline{E}$ if the zero input responses of the perturbed and unperturbed systems are equal for all $t \geq 0$ and for all initial conditions $\underline{x}(0)$.

Next consider the time domain sensitivity functions defined by the Eq.

$$\underline{w}^{i}(t,\underline{\alpha}) \triangleq \frac{\partial \underline{y}}{\partial \alpha_p}(t,\underline{\alpha}) \qquad (7)$$

where α_p is the p^{th} component of the parameter r-vector $\underline{\alpha}$ and $\underline{y}$ is the output q-vector.

Ideally, the time domain sensitivity functions defined in eq. (7) should vanish identically for all $\underline{x}(0)$ and t. For differentially small parameter changes in each α_p the difference between the zero input responses of the perturbed and unperturbed system is just

$$\sum_{p=1}^{r} \frac{\partial \underline{y}}{\partial \alpha_p} d\,\alpha_p .$$

Assuming that the differentials $d\alpha_p$ are independent "total parameter invariance" for 'small' but arbitrary parameter changes is equivalent to identically zero-sensitivity functions and an invariant response with respect to α_p implies

$$\frac{\partial \underline{y}}{\partial \alpha_p} = \underline{0} \qquad (8)$$

where

$$\alpha_p \in \underline{\alpha} \text{ for all } t \text{ and all } \underline{x}(0).$$

For the case of $\underline{v} = \underline{0}$ in Eq. (3) the following have been established [2].

Theorem

For an invariant response of a free, linear, time-invariant, multivariable system subjected to parameter variations that are arbitrary and independent in nature, it is necessary that the system be unobservable.

Theorem

A necessary and sufficient condition that the zero-input response $\underline{y}(t)$ of the system (3) be totally parameter invariant with respect to parameter variations that are arbitrary and independent in nature is that

$$[c]\,[z] = [0] \qquad (9)$$

where

$$[z] = [e^{\lambda_1 t}\left\{(\zeta_1\theta_1 + \beta_1)\ \underline{r}_1 + \theta_1 \frac{\partial \underline{r}_1}{\partial \alpha_p}\right\} + \dots$$

$$\dots + e^{\lambda_n t}\left\{(\zeta_n\theta_n + \beta_n)\ \underline{r}_n + \theta_n \frac{\partial \underline{r}_n}{\partial \alpha_p}\right\}]$$

where

$$\beta_i = < \frac{\partial \underline{\ell}_i}{\partial \alpha_p}, \underline{x}(0) > \tag{10}$$

$$\theta_i = < \underline{\ell}_i, \underline{x}(0) >$$

$$\zeta_i = t \frac{\partial \underline{\ell}_i}{\partial \alpha_p}$$

ℓ_i = the $i^{\underline{th}}$ eigenrow corresponding to the $i^{\underline{th}}$ eigenvalue λ_i

$\underline{r}_i$ = the $i^{\underline{th}}$ eigenvector corresponding to the $i^{\underline{th}}$ eigenvalue λ_i

$\frac{\partial \lambda_i}{\partial \alpha_p}$ = the $i^{\underline{th}}$ eigenvalue sensitivity function

$\frac{\partial \underline{\ell}_i}{\partial \alpha_p}$ = the $i^{\underline{th}}$ left hand (eigenrow) eigenvector sensitivity function

$\frac{\partial \underline{r}_i}{\partial \alpha_p}$ = the $i^{\underline{th}}$ right hand eigenvector sensitivity function

and $< \underline{\ell}_i, r_j > = \delta_{ij}$, the Kronecker delta

for

$i = 1, 2, \ldots, n.$

CIRCULANT MATRICES

A class of matrices, circulants, which have the inherent property that their eigenvectors are invariant with respect to arbitrary parameter variations will now be introduced [3].

Definition: A circulant is an n x n matrix of the form

$$\underline{S} = \begin{bmatrix} s_1 & s_2 & s_3 & \cdots & s_n \\ s_n & s_1 & s_2 & \cdots & s_{n-1} \\ \vdots & & & & \\ s_2 & s_3 & s_4 & \cdots & s_1 \end{bmatrix} \tag{11}$$

Next define the permutation matrix $\underline{P}$.

$$\underline{P} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{12}$$

Then

(i) $$\underline{S} = \sum_{j=1}^{n} s_j \underline{P}^{j-1} \qquad (13)$$

(ii) The eigenvalues of $\underline{S}$ are $\lambda_k = \sum_{j=1}^{n} s_j h_k^{j-1}$ (14)

where $$h_k = \cos\left(\frac{2\pi k}{n}\right) + i \sin\left(\frac{2\pi k}{n}\right)$$

(iii) All circulants have as eigenvectors

$$R_k = n^{-1/2} \left(h_k^{n-1}, h_k^{n-2}, \ldots, h_k, 1\right) \qquad (15)$$

$$k = 1, 2, \ldots, n$$

From the above discussion, it is obvious that if the $\underline{A}$ matrix of eq. (3) is a circulant and if the system were subjected to parameter variations, then to realize either a totally parameter invariant response or ε-invariant response (meaning invariance up to degree ε) one should consider only the eigenvalue sensitivity functions of the system. In the above statement it is assumed, of course, that the parameter variations are small so that the matrix $\underline{A}$ is still a circulant after the change in the nominal value of the parameters.

Example Of A System With An Invariant System Response

Consider the multivariable system defined by the set of matrix equations given below

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} a_1(\underline{\alpha}) & a_2(\underline{\alpha}) & a_3(\underline{\alpha}) \\ a_3(\underline{\alpha}) & a_1(\underline{\alpha}) & a_2(\underline{\alpha}) \\ a_2(\underline{\alpha}) & a_3(\underline{\alpha}) & a_1(\underline{\alpha}) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad (16)$$

$$\underline{y} = [1 \quad 1 \quad 1] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \qquad (17)$$

Let $\alpha_p \; \varepsilon \; (\alpha_1, \alpha_2, \ldots, \alpha_p \ldots, \alpha_r)$.

Since the $\underline{A}$ matrix of Eq. (16) is a circulant, a sufficient condition for a parameter invariant response is that

$$\frac{\partial \lambda_i}{\partial \alpha_p} = 0 \quad \text{for } i = 1, 2, 3. \qquad (18)$$

The eigenvalue sensitivity functions of the system of Eq. (16) (also see figure 1) are

$$\frac{\partial \lambda_1}{\partial \alpha_p} = \frac{\partial a_1}{\partial \alpha_p} + \frac{\partial a_2}{\partial \alpha_p} + \frac{\partial a_3}{\partial \alpha_p} \tag{19-a}$$

$$\frac{\partial \lambda_2}{\partial \alpha_p} = \frac{\partial a_1}{\partial \alpha_p} + \omega \frac{\partial a_2}{\partial \alpha_p} + \omega^2 \frac{\partial a_3}{\partial \alpha_p} \tag{19-b}$$

$$\frac{\partial \lambda_3}{\partial \alpha_p} = \frac{\partial a_1}{\partial \alpha_p} + \omega^2 \frac{\partial a_2}{\partial \alpha_p} + \omega \frac{\partial a_3}{\partial \alpha_p} \tag{19-c}$$

Since the eigenrow and eigenvector sensitivity functions are zero, the sensitivity of the system response [2] is given by the equation,

$$\frac{\partial \underline{y}}{\partial \alpha_p} = [c]\,[t \sum_{i=1}^{3} e^{\lambda_i t} \frac{\partial \lambda_i}{\partial t} < \underline{\ell}_i, \underline{x}(0) > \underline{r}_i] \tag{20}$$

where $\underline{\ell}_i$ and $\underline{r}_i$ are the normalized i^{th} left hand and right hand eigenvectors of the system.

Let the parameter α_p vary such that

$$\begin{aligned} a_1 &\to a_1 + \delta \\ a_2 &\to a_2 - \delta/2 \\ a_3 &\to a_3 - \delta/2 \end{aligned} \tag{21}$$

From eq. (19-a) it follows that $\frac{\partial \lambda_1}{\partial \alpha_p} = 0$.

Also $[c][\underline{r}_1] = [c][\underline{r}_2] = \underline{0}$

Hence the conclusion from the above is that

$$\frac{\partial \underline{y}}{\partial \alpha_p} = \underline{0} \tag{22}$$

for the parameter variation, indicated. It will be interesting to check the above result using Cruz and Perkins criterion of invariance [2], which states that:

"A necessary and sufficient condition that the zero-input response of the system (3) be totally parameter invariant with respect to a perturbation $\underline{D}$ is

$$\underline{C}\,\underline{A}^k\,\underline{D} = \underline{0} \tag{23}$$

for k = 0, 1, 2, ..., n-1."

Thus

$$\underline{C}\,\underline{A}^o\,\underline{D} = [1\ 1\ 1] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta & -\delta/2 & -\delta/2 \\ -\delta/2 & \delta & -\delta/2 \\ -\delta/2 & -\delta/2 & \delta \end{bmatrix} = 0 \tag{24}$$

and

$$\underline{C}\ \underline{A}\ \underline{D} = [1\ 1\ 1]\begin{bmatrix} a_1 & a_2 & a_3 \\ a_3 & a_1 & a_2 \\ a_2 & a_3 & a_1 \end{bmatrix}\begin{bmatrix} \delta & -\delta/2 & -\delta/2 \\ -\delta/2 & \delta & -\delta/2 \\ -\delta/2 & -\delta/2 & \delta \end{bmatrix} = 0 \tag{25}$$

Similarly it can be proved that

$$\underline{C}\ \underline{A}^2\ \underline{D} = 0 \tag{26}$$

A significant difference between the Cruz and Perkins criterion of invariance and our criterion is that in the former case, the perturbation matrix $\underline{D}$ has to be explicitly known while in the latter the matrix $\frac{\partial \underline{A}}{\partial \alpha_p}$ has to be first determined [2].

Generally speaking the perturbation matrix $\underline{D}$ is not available and the recourse to the second criterion may prove more useful.

Comments

It is of interest to note that when the $\underline{A}$ matrix of Eq. (3) is a circulant such that

$$\underline{A} = \sum_{j=1}^{n} a_j\, \underline{P}^{j-1} \tag{27}$$

where $\underline{P}$ is a permutation matrix defined in eq. (12) and the parameter variations are such that

$$\frac{\partial a_1}{\partial \alpha_p} = \frac{\partial a_2}{\partial \alpha_p} = \ . \ . \ . = \frac{\partial a_n}{\partial \alpha_p} \tag{28}$$

then the eigenvalue sensitivity function of all the complex eigenvalues vanishes identically.

CONCLUSIONS

The problem of invariance of system response of linear, time invariant, multivariable systems was examined from a new point of view. The notions of eigensystem sensitivity, unobservability and invariance were shown to be related via the time domain sensitivity functions of the system response. An inherent property of invariance of the eigenvector (both left hand and right hand) of a certain class of matrices-circulants - was discussed. An example of a system with an invariant response for certain parameter variations was also given. Generally speaking, in practice it is unreasonable to expect the sensitivity functions to be reduced to zero and it would be more realistic to reduce some bounds of these functions. Again the bounds used should be reasonably sharp and should have some physical significance. Although this is a rather difficult problem, attempts are being made to obtain some such bounds which will eventually be useful in reducing the system-response sensitivity.

REFERENCES

1. Cruz, J. B. Jr., and Perkins, W. R., "On Invariance and Sensitivity," IEEE Convention Record, New York, 1966.
2. Reddy, D. C., "Parameter Sensitivity of Linear, Time-invariant, Dynamic Systems," Report R-315, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois.
3. Hamburger, H. L. and Grimshaw, M. E., "Linear Transformations in n-dimensional Vector Space," University Press, Cambridge.

Example of a System With an Invariant Response

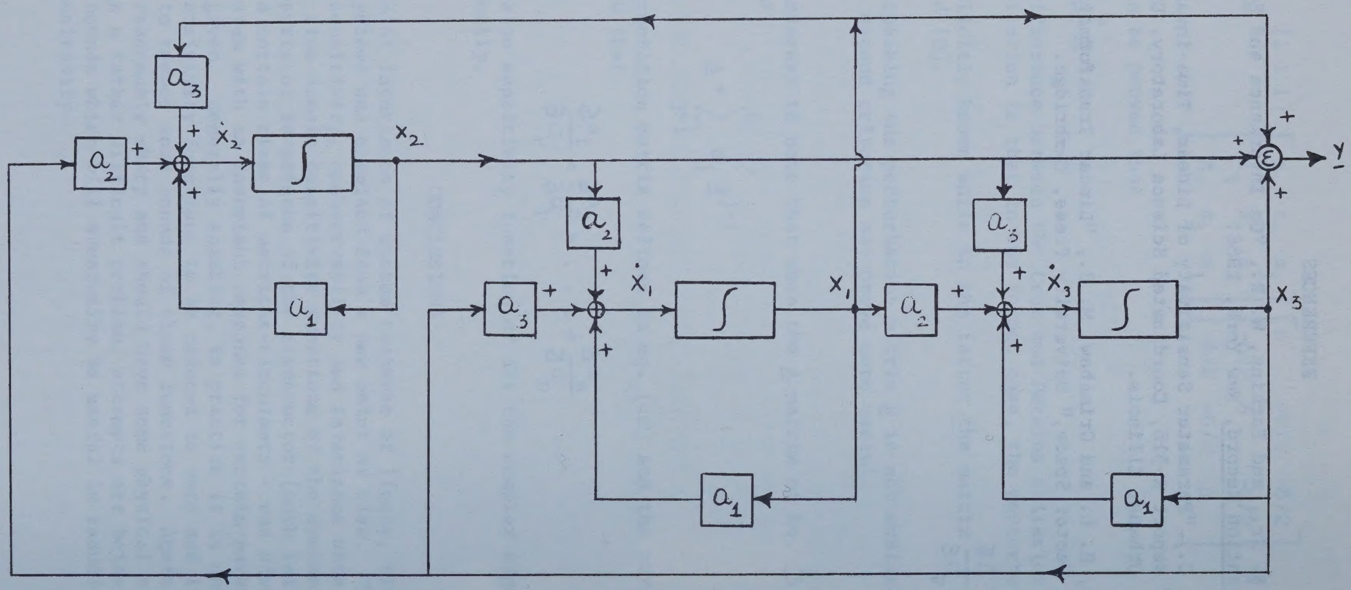

a_3
a_2
$\dot{x}_2$
$\int$
x_2
a_1
a_2
a_3
$\dot{x}_1$
$\int$
x_1
a_1
a_2
a_3
$\dot{x}_3$
$\int$
x_3
a_1
ε
$\underline{y}$
+

Fig. 1

CHARACTERIZATION OF A SIMPLE NONLINEAR SYSTEM

M. L. LIOU
Bell Telephone Laboratories, Inc.
No. Andover, Massachusetts

ABSTRACT - Characterization techniques of a simple nonlinear system are described in this paper. The nonlinear system is composed of a third order polynomial zero-memory nonlinear device and a linear time-invariant system connected in tandem. The input signal of the nonlinear system is a family of zero mean Gaussian noise with specified power-density spectrum. Based on a set of measured data, that is, the power-density spectrum of the output and the cross-power-density spectrum of the input and output, we can determine the characteristic of the zero-memory nonlinear device as well as the transfer function of the linear time-invariant system. A possible application of these techniques is to characterize a repeater section in a broadband transmission system.

INTRODUCTION

Consider a simple nonlinear system shown in Fig. 1, where the input-output relationship of the zero-memory nonlinear device is

$$z(t) = x(t) + bx^2(t) + cx^3(t) \qquad (1)$$

and $H(\omega)$ is the transfer function of a linear system. The input $x(t)$ is Gaussian with zero mean and its power-density spectrum is expressed as

$$S_x(\omega) = k\Phi(\omega) \qquad (2)$$

where k is an arbitrary constant and $\Phi(\omega)$ is a specified frequency function.

In this paper, we present some characterization techniques to determine the coefficients of the zero-memory nonlinear device, b and c, as well as the transfer function of the linear system, $H(\omega)$, based on a set of measured data, that is, the power-density spectrum of the output, $S_y(\omega)$, and the cross-power-density spectrum of the input and output, $S_{xy}(\omega)$.

A possible application of these techniques is to characterize a repeater section in a broadband transmission system in connection with the use of complementary nonlinear equalization [1-4]. Referring to Fig. 2, a repeater section with small nonlinearity can be represented by a third order polynomial zero-memory nonlinear device N and some delay and amplitude distortion, $H(\omega)$, in cascade. In a first step of equalization, the amplitude and delay distortion is eliminated in the (complementary) amplitude and delay equalizers, $1/H(\omega)$. A complementary nonlinear equalizer which possesses the inverse characteristic of N then cancels the remaining nonlinear distortion. Obviously, a successful equalization depends heavily on the accurate characterization of N and $H(\omega)$.

The characterization techniques presented in this paper are not as general as the orthogonal functional series [5]. However, they are simple and extremely useful when the nonlinearity of a given physical system is small.

ANALYSIS

Let $R_x(\tau)$ and $R_z(\tau)$ be respectively the autocorrelation functions of x(t) and z(t). Since x(t) has zero mean, it can be shown that [6]

$$R_z(\tau)=b^2R_x{}^2(o)+ [1+3c\,R_x(o)]^2\,R_x(\tau)+2b^2R_x{}^2(\tau)+6c^2R_x{}^3(\tau) \quad (3)$$

Let $S_y(\omega)$ and $S_z(\omega)$ be respectively the power-density spectra of y(t) and z(t), then

$$S_y(\omega) = |H(\omega)|^2\, S_z(\omega)$$

(Equation Cont'd)

$$= |H(\omega)|^2 \left\{ b^2 R_x{}^2(o)\delta(\omega) + [1+3c\ R_x(o)]^2\ S_x(\omega) + 2b^2 S_x(\omega) * S_x(\omega) + 6c^2 S_x(\omega) * S_x(\omega) * S_x(\omega) \right\} \quad (4)$$

where $\delta(\omega)$ is the frequency delta function and * denotes the convolution in the frequency domain.

It can also be shown that the cross-correlation function of x(t) and y(t) is

$$R_{xy}(\tau) = [1+3cR_x(o)] \int h(\alpha) R_x(\tau-\alpha)d\alpha$$

where $h(\alpha)$ is the impulse response of the linear system. Hence, the cross-power-density spectrum of x(t) and y(t) is

$$S_{xy}(\omega) = [1+3cR_x(o)]\ H(\omega) S_x(\omega) \quad (5)$$

The characterization of the nonlinear system in Fig. 1 can be easily accomplished using (4) and (5).

CHARACTERIZATION PROCEDURE

Let $\Phi(\omega)$ be invariant. We change the input power-density spectrum $S_x(\omega)$ by adjusting the constant k. Substituting (2) into (5) we obtain

$$S_{xy}(\omega) = H(\omega)\Phi(\omega)k + 3cdH(\omega)\Phi(\omega)k^2$$

where

$$d = \frac{1}{2\pi} \int \Phi(\omega)d\omega$$

When k varies from k_1, k_2 ..., to k_N, we have the following matrix equation

$$K\ \underline{u}(\omega) = \underline{S}_{xy}(\omega) \quad (6)$$

where

$$K = \begin{bmatrix} k_1 & k_1^2 \\ k_2 & k_2^2 \\ \vdots & \\ k_N & k_N^2 \end{bmatrix}$$

$$\underline{u}(\omega) = \begin{bmatrix} u_1(\omega) \\ u_2(\omega) \end{bmatrix} = \begin{bmatrix} H(\omega)\Phi(\omega) \\ 3cd\ H(\omega)\Phi(\omega) \end{bmatrix}$$

$$\underline{S}_{xy}(\omega) = [S_{xy1}(\omega) \quad S_{xy2}(\omega) \ \ldots \ S_{xyN}(\omega)]^T$$

The numeral subscript associated with $S_{xy}(\omega)$ indicates the cross-power-density spectrum measured due to the corresponding value of k and the superscript T denotes the transposition.

In principle, only two values of k are sufficient to solve for (6). However, if the model in Fig. 1 is an approximate representation of an actual system and the measurements are in error, more values of k should be used in order to represent the actual system closely over the given domain of k.

(6) can be solved in the least squares sense [7]

$$\underline{u}(\omega) = (K^T K)^{-1} K^T \underline{S}_{xy}(\omega)$$

Thus, we have

$$H(\omega) = \frac{u_1(\omega)}{\Phi(\omega)}$$

Let

$$c_i = \frac{u_2(\omega_i)}{3d\ u_1(\omega_i)}$$

Where ω_i is an arbitrary angular frequency inside the band of interest. The c_i's may not be the same for different values of ω_i as a result of measurement error, approximation, etc. Hence, the coefficient c of the zero-memory nonlinear device is determined by taking the average value of c_i

$$c = \frac{1}{M} \sum_{i=1}^{M} c_i$$

From (4), we neglect the dc power and let

$$b_i(k) = \left\{ \frac{S_{yk}(\omega)/|H(\omega)|^2 - 6c^2 S_x(\omega)*S_x(\omega)*S_x(\omega) - [1+3cR_x(o)]^2 S_x(\omega)}{2S_x(\omega)*S_x(\omega)} \right\}^{\frac{1}{2}}_{\omega=\omega_i}$$

where $S_{yk}(\omega)$ is the measured power-density spectrum at the output corresponding to $S_x(\omega) = k\Phi(\omega)$. Again, the coefficient b of the zero-memory nonlinear device due to a particular value of k can be determined as

$$b(k) = \frac{1}{M} \sum_{i=1}^{M} b_i(k)$$

For given input domain $k = k_1, k_2, \ldots, k_N$, the coefficient b can be determined as the average value of b(k)

$$b = \frac{1}{N} [b(k_1) + b(k_2) + \ldots + b(k_N)]$$

An alternative method of characterization when the nonlinearity is small is as follows. From (5) choose a small value of k, then

$$S_{xy}(\omega) \cong H(\omega)\ S_x(\omega)$$

Hence

$$H(\omega) = \frac{S_{xy}(\omega)}{S_x(\omega)}$$

We cut a narrow slot in $S_x(\omega)$ by using a band-rejection filter and measure the power-density spectrum in this slot $(\omega \neq 0)$ at the output of the nonlinear system. (4) becomes

$$S_y(\omega) = 2k^2 |H(\omega)|^2 \Phi(\omega)*\Phi(\omega) b^2 + 6k^3 |H(\omega)|^2 \Phi(\omega)*\Phi(\omega)*\Phi(\omega) c^2$$

Let $k = k_1, k_2, \ldots, k_N$, we have the matrix equation

$$K\, \underline{v}(\omega) = \underline{S}_y(\omega) \qquad (7)$$

where

$$K = \begin{bmatrix} k_1^{\ 2} & k_1^{\ 3} \\ k_2^{\ 2} & k_2^{\ 3} \\ \vdots & \\ k_N^{\ 2} & k_N^{\ 3} \end{bmatrix}$$

$$\underline{v}(\omega) = \begin{bmatrix} v_1\ (\omega) \\ v_2\ (\omega) \end{bmatrix} = \begin{bmatrix} 2|H(\omega)|^2 \Phi(\omega)*\Phi(\omega) b^2 \\ 6|H(\omega)|^2 \Phi(\omega)*\Phi(\omega)*\Phi(\omega)\ c^2 \end{bmatrix}$$

$$\underline{S}_y\ (\omega) = [S_{y1}(\omega)\ \ S_{y2}(\omega)\ \ \ldots\ \ S_{yN}(\omega)]^T$$

The numeral subscript associated with $S_y(\omega)$ indicates the power-density spectrum measured at the output due to the corresponding value of k.

Again (7) can be solved in the least squares sense

$$\underline{v}(\omega) = (K^T K)^{-1} K^T \underline{S}_y(\omega)$$

Let

$$b_i = \left[\frac{v_1(\omega)}{2\,|H(\omega)|^2 \Phi(\omega)*\Phi(\omega)} \right]^{\frac{1}{2}}_{\omega=\omega_i}$$

$$c_i = \left[\frac{v_2(\omega)}{6\,|H(\omega)|^2 \Phi(\omega)*\Phi(\omega)*\Phi(\omega)} \right]^{\frac{1}{2}}_{\omega=\omega_i}$$

Then the coefficients b and c are determined as

$$b = \frac{1}{M} \sum_{i=1}^{M} b_i$$

$$c = \frac{1}{M} \sum_{i=1}^{M} c_i$$

ACKNOWLEDGMENTS

The author wishes to thank A. J. Giger for providing the references 2, 3, and 4. He is also grateful to J. J. Lang and F. J. Witt for many valuable suggestions.

REFERENCES

1. A. J. Giger, Bell Telephone Laboratories, Inc., Private Communication.

2. G. W. Holbrook and G. N. Henderson, "Minimizing Cross Modulation in Multichannel Amplifiers by Complementary Distortion", Electronics and Communications, November 1963, pp. 28-30.

3. R. Legros and R. Besnard, "Anti-distortion Circuit for Tube and Transistor Amplifiers", L'Onde Electrique, September 1965, pp. 1081-1092.

4. A. Knoll, "A Low Distortion Frequency Modulator with Varactor Diodes and Logarithmic Amplifier", Nachrichtentechnische Zeitschift (NTZ), August 1965, pp. 455-463.

5. Y. W. Lee and M. Schetzen, "Measurement of the Wiener Kernels of a Nonlinear System by Cross-correlation", International Journal of Control, September 1965, pp. 237-254.

6. M. L. Liou, "Noise in an FM System due to an Imperfect Linear Transducer", to be published in B.S.T.J., November, 1966.

7. R. Penrose, "A Generalized Inverse for Matrices", Proc. Cambridge Philos. Soc., 51(1955), pp. 406-413.

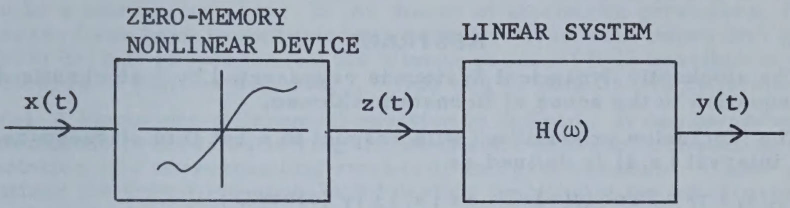

ZERO-MEMORY
NONLINEAR DEVICE
LINEAR SYSTEM
x(t)
z(t)
y(t)
H(ω)

Fig. 1 - Block Diagram of a Simple Nonlinear System

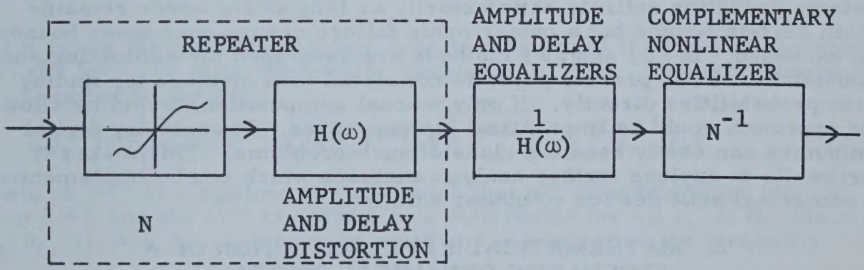

REPEATER
AMPLITUDE
AND DELAY
EQUALIZERS
COMPLEMENTARY
NONLINEAR
EQUALIZER
H(ω)
$\frac{1}{H(\omega)}$
N^{-1}
N
AMPLITUDE
AND DELAY
DISTORTION

Fig. 2 - Complementary Nonlinear Equalization

ON THE INCLUSION PROBABILITY OF A STOCHASTIC DYNAMICAL SYSTEM*)

LEO J. VAN MELLAERT
Department of Electrical Engineering
Polytechnic Institute of Brooklyn
Brooklyn, New York

ABSTRACT

The stochastic dynamical system is represented by a stochastic differential equation in the sense of Bernstein-Gihman.

The "inclusion probability" with respect to a set D in state-space and a time interval $[s, t]$ is defined as

$$P_D(x, s;t) = \text{Prob}\{y(\tau)\varepsilon D: \tau \varepsilon [s,t] \mid y(s)=x\varepsilon D\}$$

wherein $y(\tau)$ is the state of the system at time τ.

It can be shown that under suitable conditions, the inclusion probability satisfies Kolmogorov's backward equation. A simplified proof will be given here.

A practical example is given for which an analytic closed form solution could be obtained. More complicated problems can be solved on the digital computer using numerical methods for solving parabolic partial differential equations.

1. INTRODUCTION

Stochastic dynamical systems have been studied frequently on the basis of "white noise" models. This paper will assume that a valid model for the physical system exists. Some of the connections between the physical system and its model are discussed in [1].

As Kushner points out in [2], must of the past work in stochastic control theory deals with moments of the state-variables. However some stochastic systems can be evaluated more realistically based on the probabilities of desirable or undesirable events. For example, one can imagine systems operating entirely satisfactorily as long as the error remains within certain bounds but a catastrophic failure occurs once these bounds are exceeded. In [2] and [3] methods are developed for estimating such probabilities. The present paper is connected with methods for finding these probabilities directly. If only manual computations would be allowed this approach would be impractical in most cases. Present day digital computers can easily handle a class of such problems. This makes it worthwhile to explore further analytic methods which can be implemented by numerical schemes for computer solution.

2. MATHEMATICAL REPRESENTATION OF A STOCHASTIC DYNAMICAL SYSTEM.

The state of a deterministic m-dimensional dynamical system can be represented by a point $y=(y_1, y_2, \ldots, y_m)$ in the m-dimensional Euclidean space E^m. Given that the state is at a point $y(s)$ at the initial time $t=s$, the

---

*) This paper contains some results of the research for a doctoral dissertation at the Polytechnic Institute of Brooklyn. The research was sponsored under National Science Foundation Grant no. GK-527.

trajectory of the state point for all time $t > s$ is given by the solution of the system-equations

$$dy_i = f_i (y, t)\, dt \qquad (i = 1, 2, \ldots, m). \tag{2.1}$$

The state of a stochastic m-dimensional dynamical system is a stochastic process $y(t, \omega) : y \varepsilon E^m$, $\omega \varepsilon \Omega$, Ω being the sample space. In most cases of interest for engineering applications the state of the system turns out to be a Markov process. In the theory of stochastic processes, Markov processes have been investigated intensively. It is well known that going out from certain processes (as the Wiener process) it is possible to derive new classes of Markov processes. This can be done in two different ways:

(a) A stochastic differential equation is defined. It can be shown that the stochastic process satisfying the stochastic differential equation is also the solution of a corresponding stochastic integral equation. Under suitable conditions the transition probability density function of the Markov process being the solution of the above equations will satisfy Kolmogorov's equations [4], [5];

(b) A second approach due to Bernstein seems to be more natural for engineering applications, since it establishes a direct link between an incremental equation of the stochastic dynamical system and the transition probability density function, which is often of prime interest in practical problems [6], [7]. The relationship between this approach and the previous one is studied in [8]. This paper will follow the formulation of [6]. The results needed in the subsequent discussion are summarized next.

Consider a time interval $[s, t]$ composed of n non-overlapping subintervals of length $\Delta t^{(k)}$ $(k=1, 2, \ldots, n)$ such that

$$t = s + \sum_{k=1}^{n} \Delta t^{(k)} . \tag{2.2}$$

Let $y(s)$ be a random variable given at the initial time s. The random variable $y^{(n)}(t)$ at time t is defined as

$$y_i^{(n)}(t) = y_i(s) + \sum_{k=1}^{n} \Delta y_i^{(k)} \quad (i=1, 2, \ldots, m). \tag{2.3}$$

Let $p_o(y)$ be the joint distribution density function of the random variable $y(s)$ and let $p_n(y, t)$ be the density function of $y^{(n)}(t)$. The increments Δy_i are assumed to satisfy

$$\Delta y_i^{(k)} = b_i(y, t)\Delta t^{(k)} + c_i(y, t, \alpha^{(k)}) \sqrt{\Delta t^{(k)}} \quad (i=1, 2, \ldots, m) \tag{2.4}$$

in which $\alpha^{(k)}$ is a random variable such that $\alpha^{(j)}$ is independent from $\alpha^{(k)}$ when $j \neq k$ and the $\alpha^{(k)}$ are identically distributed for all k. If the functions $\partial c_i / \partial y_i$ $(i, j = 1, 2, \ldots, m)$ are bounded and if b satisfies the inequality

$$\sum_{i=1}^{m} y_i b_i < K \left(1 + \sum_{i=1}^{m} y_i^2\right), \quad K > 0 \tag{2.5}$$

then the joint distribution density function of the random variable $y^{(n)}(t)$ will tend to a limit function $p(y, t)$ when $n \to \infty$ such that $\max(\Delta t^{(k)}) \to 0$. If in addition the functions b_i are analytic and if the initial distribution $p_o(y)$ has second order partial derivatives then the limit function $p(y, t)$ satisfies the Fokker-Planck or Kolmogorov's forward equation:

$$\frac{\partial p}{\partial t} = \sum_{i,j=1}^{m} \frac{\partial^2 (a_{ij} p)}{\partial y_i \, \partial y_j} - \sum_{i=1}^{m} \frac{(b_i p)}{y_i} \tag{2.6}$$

wherein

$$a_{ij} = \tfrac{1}{2} E_\alpha \{c_i c_j\} \quad , \quad E_\alpha \{c_i\} = 0 \quad ,$$

E_α is denoting the expectation with respect to the random variable α. The initial condition for eqn. (2.6) is

$$p(y, s) = p_o(y) .$$

It has been shown that a stochastic process exists which is the limit of the sequence of stochastic processes defined in (2.3) when $n \to \infty$ [9]. In this last reference less restrictive conditions are imposed on the random variables α.

If the initial distribution density function p_o tends to an impulse function, the state at the initial time s will be deterministic $y(s)=x$ and the function $p(y, t)$ will become the transition probability density function $p(x, s ; y, t)$ in which the dependence upon the parameters x, s is indicated explicitely. $p(x, s; y, t)$ satisfies the Fokker-Planck equation (2.6) in its y, t arguments and also satisfies Kolmogorov's backward equation [10]:

$$\frac{\partial p}{\partial s} = - \sum_{i,j=1}^{m} a_{ij}(x, s) \frac{\partial^2 p}{\partial x_i \, \partial x_j} - \sum_{i=1}^{m} b_i(x, s) \frac{\partial p}{\partial x_i} . \tag{2.7}$$

3. INCLUSION PROBABILITY OF A MARKOV PROCESS.

Consider a stochastic dynamical system the state of which is a Markov process in the state-space E^m. In many problems it is of interest to know whether the sample trajectories of the process will be restricted to a given subset D of E^m. Given that at the initial time s the state is at the point $x \varepsilon \subset E^m$, the "Inclusion Probability" P_D is defined as the conditional probability

$$P_D(x, s;t) = \mathrm{Prob}\,\{y(\tau) \varepsilon D \cdot \ \tau \varepsilon [s,t] \ | y(s) = x \varepsilon D\} \tag{3.1}$$

in which $y(\tau)$ represents the state of the system at time τ.

It can be shown that this inclusion probability satisfies the partial differential equation (2.7), but with different initial and boundary conditions. This property is pointed out without a complete proof in [11] and [12]. A rigorous proof is given in [5] but rests upon a considerable amount of preceding work and presupposes a sufficient training in functional analysis and measure theory on the part of the reader.

Define the density function p_D such that

$$p_D(x, s;y, t) = \mathrm{Prob}\{ y(t) \varepsilon \, dy \subset D, y(\tau) \varepsilon D: \ \tau \varepsilon [s, t] \ | y(s) = x \varepsilon D\} ,$$

wherein dy represents a neighborhood of the point y. For continuous Markov processes the following assymptotic relationship holds

$$p(x, t;y, t+\tau) = p_D(x, t;y, t+\tau) + 0(\tau). \tag{3.2}$$

By means of this relationship it is easy to prove the following theorem:

> If a Markov process has a transition probability density function $p(x, s;y, t)$ which satisfies the Fokker-Planck equation (2.6), then the inclusion probability in a set D: $P_D(x, s;t)$, satisfies Kolmogorov's backward equation (2.7).

Proof. Because of the Markov property

$$p_D (x, s-\sigma; y, t) = \int_D p_D(x, s-\sigma; \eta , s)\, p_D (\eta , s; y, t) d\eta .$$

Using the assymptotic relationship (3. 2)

$$p_D(x, s-\sigma; y, t) = \int_D p(x, s-\sigma; \eta , s) p_D(\eta , s; y , t) d\eta + 0(\sigma). \tag{3. 3}$$

For $s \to t$, p behaves like a delta function: $\delta (x-y)$, but for all $s<t$ p is continuous and has a first order partial derivative with respect to s. Expanding p in (3. 3) about $(s-\sigma_o)$: $0 < \sigma_o < \sigma$,

$$p_D(x, s-\sigma; y, t) = \int_D p(x, s-\sigma_o; \eta , s) p_D(\eta , s; y, t)\, d\eta$$

$$-(\sigma-\sigma_o) \int_D \frac{\partial}{\partial s} p\, (x, s-\sigma' ; \eta , s)\, p_D(\eta , s; y, t) d\eta + 0(\sigma)$$

wherein $\sigma' \varepsilon [\sigma_o, \sigma]$. Using (2. 7) and letting $\sigma_o \to 0$

$$p_D(x, s-\sigma; y, t) - p_D(x, s; y, t)$$

$$= \sigma\Big[\sum_{i, j\;\; 1}^{m} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i\;\; 1}^{m} b_i \frac{\partial}{\partial x_i} \Big] \int_D p(x, s-\sigma' ; \eta , s) p_D(\eta , s; y, t) d\eta + 0(\sigma).$$

Dividing by σ and letting $\sigma \to 0$ it follows that p_D satisfies eqn. (2. 7). The inclusion probability P_D is obtained by integrating $p_p(x, s; y, t)$ with respect to y over the set D. Thus integrating eqn. (2. 7) satisfied by p_D with respect to y finishes the proof of the theorem:

$$\frac{\partial}{\partial s} P_D(x, s; t) = - \sum_{i, j=1}^{m} a_{ij}(x, s) \frac{\partial^2 P_D}{\partial x_i \partial x_j} - \sum_{i=1}^{m} b_i (x, s) \frac{\partial P_D}{\partial x_i}. \tag{3. 4}$$

From the definition (3. 1) of P_D it follows that for every interior point x of the set D

$$\lim_{s \to t} P_D (x, s; t) = 1. \tag{3. 5}$$

This relation establishes the initial condition for (3. 4)*). If D is an open set, then certainly

$$P_D(x, s; t) = 0 \quad x \varepsilon\, \partial D \tag{3. 6}$$

in which ∂D represents the boundary of D. It can be shown that for continuous Markov processes the first exit time from the interior of a set coincides with the first exit time from that set [5]. The boundary condition (3. 6) is thus valid independently of the fact whether D is an open or a closed set. (See also the heuristic explanation in [12]).

4. NUMERICAL COMPUTATION OF THE INCLUSION PROBABILITY

Even in relatively simple cases of interest an analytic solution of eqn. (3. 4) appears to be difficult to obtain. Some authors have derived bounds for the inclusion probability [2],[3]. These methods seem however to be limited to simple dynamical systems and no a priori information has been obtained on the closeness of these bounds to the actual values of the inclusion probability.

---

*) "Final" condition may be a better term since eqn. (3. 4) has to be solved for $s \leq t$.

An alternate approach is to solve equation (3.4) numerically. A large class of problems can be accommodated efficiently on an average size digital computer. An advantage of the numerical solution is that often the exact solution can be approximated as close as desired, provided that the number of arithmetic operations is allowed to increase as necessary. Which of these two approaches is to be preferred depends on the particular problem.

From the previous sections it is recalled that the inclusion probability of a dynamical system characterized by the eqns. (2.4), satisfies the eqn. (3.4) subject to the initial condition (3.5) and the boundary condition (3.6). Eqn. (3.4) is a second order parabolic partial differential equation. For such equations it is customary to introduce a new time variable so that the initial condition is specified at time 0 and that the equation is solved for positive values of time. Introducing

$$u(x, t) = P_D(x, -t; 0), \tag{4.1}$$

the partial differential equation (3.4) becomes

$$\frac{\partial}{\partial t} u(x, t) = \sum_{i,j=1}^{m} a_{ij}(x, t) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{m} b_i(x, t) \frac{\partial u}{\partial x_i} \quad *) \tag{4.2}$$

with the initial and boundary conditions

$$\lim_{t \to 0} u(x, t) = 1 \tag{4.3}$$

$$u(x, t) = 0 \quad x \varepsilon \partial D. \tag{4.4}$$

The numerical method for solving eqn. (4.2) is discussed by A.A. Smarskii [13]. A rather condensed description of this method is given here. The more interested reader is referred to the references given for specific information on the validity and the convergence of the method.

For simplicity it is assumed that the matrix (a_{ij}) is diagonal and that the function a_{ij} does not depend upon the space variable x_i. Most cases of interest will satisfy these assumptions. It should be noted that the method also applies to more general equations [13].

Equation (4.2) is split into m one-(space)dimensional partial differential equations

$$\frac{\partial}{\partial t} u(x, t) = a_{ii}(x, t) \frac{\partial^2 u}{\partial x_i^2} + b_i(x, t) \frac{\partial u}{\partial x_i} \quad (i=1, 2, \dots, m). \tag{4.5}$$

Each of these equations is now numerically solved, the i-th equation along the direction of the x_i coordinate axis. The region D is therefore spanned by an m-dimensional net of lines. In order to avoid notational complications, the distance between two adjacent net-lines will be taken as constant h, the same in all the m coordinate-directions. This is however no limitation of the method [13]. The point of intersection of net-lines is called a node. An approximation to the solution of eqn. (4.2) will be found at discrete points in time, spaced by intervals of length τ.

---

*) For notational convenience no new symbols have been introduced for the coefficients a_{ij} and b_i. It is understood however that their time arguments have been adapted consistently with the introduction of the new variable in (4.1).

The solution proceeds as follows. The equation (4.5) with $i=1$ will be solved at $t=\tau/m$ along every net-line in the direction x_1 using the initial condition (4.3) at $t=0$ and satisfying the boundary condition (4.4) at those points where the particular net-line intersects the boundary ∂D. In general these points of intersection will not coincide with nodes of the net. Momentarily it is assumed that every point of intersection is also a node point. After solving eqn. (4.5) as many times as there are net-lines in the direction of the x_1 coordinate axis, eqn. (4.5) will be solved for $i=2$ at $t=2\tau/m$ along every net-line in the direction of the x_2-axis. The initial condition at $t=\tau/m$ for each of these solutions will be the values for u found from the preceding set of solutions along the direction x_1. The boundary conditions are analogous to those for the solutions along x_1. The entire process now repeats for $i=3, 4, \dots$ until $i=m$ and thus $t=m.\tau/m=\tau$. The values for u obtained at this instant approximate the solution of eqn. (4.2) at $t=\tau$. Then the procedure is continued with $i=1$ and $t=\tau+\tau/m$ as described above.

The basic difference equation to be solved repeatedly in the execution of this process is

$$\frac{u(t) - u(t-\tau/m)}{\tau} = a\,\frac{u^+ - 2u + u^-}{h^2} + b\,\frac{u^+ - u^-}{2h}, \tag{4.6}$$

in which $u = u(t) = u(t, x_1, x_2, \dots, x_i, \dots, x_m)$

$u^+ = u(t, x_1, x_2, \dots, x_i + h, \dots, x_m)$

$u^- = u(t, x_1, x_2, \dots, x_i - h, \dots, x_m)$.

For notational convenience introduce $F = u(t-\tau/m)$, the value of F being available from the preceding solution in the direction of x_{i-1}. Rewriting eqn. (4.6)

$$(a_{ii} + b_i\frac{h}{2})\,u^+ - (2a_{ii} + \frac{h^2}{\tau})\,u + (a_{ii} - b_i\frac{h}{2})\,u^- = -\frac{h^2}{\tau}\,F \tag{4.7}$$

This is a second order difference equation which is to be solved along each net-line subject to the boundary condition at each of the intersections of this net-line with the boundary D. Labelling the nodes along a particular net-line as $0, 1, 2, \dots, N$ such that the nodes 0 and N are the respective boundary points, the eqn. (4.7) is readily solved by introducing two auxiliary variables ξ and η satisfying

$$\xi\,(n+1) = \frac{a_{ii} + b_i\frac{h}{2}}{2a_{ii} + \frac{h^2}{\tau} - (a_{ii} - b_i\frac{h}{2})\,\xi(n)}, \quad \xi\,(1) = 0 \tag{4.8}$$

$$\eta\,(n+1) = \frac{(a_{ii} - b_i\frac{h}{2})\,\eta\,(n) + \frac{h^2}{\tau}\,F}{a_{ii} + b_i\frac{h}{2}}, \quad \eta\,(1) = 0, \tag{4.9}$$

the arguments of ξ and η referring to the label of the node. The solution of eqn. (4.7) is given by

$$u(n-1) = \xi(n)\,u(n) + \eta(n)\,, \quad u(N) = 0\,. \tag{4.10}$$

In the case that the intersection point of a net-line with ∂D does not coincide with a node, the boundary points 0 and N will be chosen such that they are nodes still contained in D. The boundary conditions of (4.8) and (4.10) are then modified in order to improve the approximation of the difference scheme [13]. If the distance between node 0 and the intersection point of the net-line with ∂D is k_0 and the corresponding distance for node N is k_N then

$$\xi(1) = k_0/h$$

$$u(N) = \frac{k_N \, \eta(N)}{h - k_N \, \xi(N)} \, .$$

As illustration two examples follow which were programmed on an IBM 7040 computer. The first example, Fig. 1, is for a first order dynamical system with an incremental equation

$$\Delta y = -0.1 \, y \, \Delta t + \Delta w, \qquad E\{(\Delta w)^2\} = 0.2 \;\; \Delta t.$$

The region D is the interval (-1, 1). Plotted is $P_D(x, o; t)$.

The second example, Fig. 2, is the inclusion probability for a second order dynamical system with the incremental equations

$$\Delta y_1 = \qquad y_2 \cdot \Delta t + \Delta w_1 \qquad\qquad E\{(\Delta w_1)^2\} = .0625 \, \Delta t$$

$$\Delta y_2 = -(4y_1 + 4y_2) \;\; \Delta t + \Delta w_2 \quad , \quad E\{(\Delta w_2)^2\} = \;\; 4.0 \, \Delta t.$$

The region D is a circle in the 2-dimensional state space centered around the origin with radius one. The figure shows the contours of constant probability $P_D((x_1, x_2), 0; t)$ for t = 0.05 seconds, (x_1, x_2) being the initial coordinates of the state point.

It should be noted that in this last example a small disturbance Δw_1 appears in the first of the two system equations. It turns out that with the method of solution as described previously, there is a lower bound to the variance of Δw_1, depending upon the net parameter h and the geometry of the region D. Lack of space prohibits a further discussion of this problem. A paper describing some practical results is planned.

APPENDIX

AN ANALYTICAL SOLUTION FOR A FIRST ORDER SYSTEM

In general, the solution of eqns. (2.6) or (3.4) subject to their respective initial and boundary conditions is not an easy task. A case is presented here, for which a closed form solution can be obtained.

Consider a first order stochastic dynamical system with an incremental equation corresponding to eqn. (2.4)

$$\Delta y = -\, a \, y \, \Delta t + \Delta w \quad , \quad E\{(\Delta w)^2\} = \sigma^2 \, \Delta t. \tag{A.1}$$

This equation is sometimes written symbolically as

$$dy/dt = -\, a \, y + \dot{w}$$

wherein the symbol $\dot{w}$ stands for "white noise". This process can be defined formally as

$$\dot{w} = \lim_{t \to 0} \; (\Delta w / \Delta t) \qquad *)$$

---

*) It is clear that this limit does not exist in a strict mathematical sense because the deviation of the increments Δw is proportional to $\sqrt{\Delta t}$.

Δw being the increments of a Wiener process [14].

For the system equation (A.1) the Fokker-Planck equation (2.6) now reads

$$\frac{\partial p}{\partial t} = \tfrac{1}{2}\sigma^2 \frac{\partial^2 p}{\partial y^2} + a\,y\,\frac{\partial p}{\partial y} + ap. \tag{A.2}$$

If in solving eqn. (A.2) it is assumed that the initial distribution of p is normal centered around x with a variance σ_o^2, it is possible to show that the solution will also be normal at later times with a variance and a mean value which are also functions of time. Both these functions can be determined by substituting a normal solution into eqn. (A.2) and satisfying the initial condition. If subsequently the variance of the initial distribution tends to zero, the transition probability density function is obtained for an initial position x at time s:

$$p(x, s;y, t) = \frac{1}{\sigma}\sqrt{\frac{a}{\eta}}\,[1-e^{-2a(t-s)}]^{-\frac{1}{2}} \exp\left\{\frac{-a[y-xe^{-a(t-s)}]^2}{\sigma^2[1-e^{-2a(t-s)}]}\right\}. \tag{A.3}$$

With this expression it is possible to construct a solution for eqn. (3.4) subject to the conditions (3.4, 5), for a case where the region D is the interval $(K, +\infty)$. It is clear that the function q satisfies the boundary condition (3.6):

$$q(x, s;y, t) = p(x, s;y, t) - p(2K-x, s;y, t)$$

while q also satisfies the linear equation (3.4). Also because of the linearity of eqn. (3.4), the integral of q with respect to y over the region D will also satisfy eqn. (3.4) while the integral still satisfies the boundary condition (3.6). It is easily verified that this integral also satisfies the initial condition (3.5). The inclusion probability is thus

$$P_D(x, s;t) = \int_K^{\infty} q(x, s;y, t)\,dy,$$

which after some manipulations reduces to

$$P_D(x, s;t) = \int_K^{[2(x-K)e^{-a(t-s)} + K]} p(x, s;y, t)\,dy.$$

ACKNOWLEDGMENT

The author wishes to thank his advising professor, Dr. P. Dorato, for suggesting the topic and for the many discussions during the course of the research.

REFERENCES

[1] H. J. KUSHNER, "On the Status of Optimal Control and Stability for Stochastic Systems", 1966 IEEE International Convention Record, Part 6, 143-151 (1966).

[2] H. J. KUSHNER, "Finite Time Stochastic Stability and the Analysis of Tracking Systems", IEEE Trans. on Automatic Control, vol. AC-11, 219-227, April 1966.

[3] B. FRIEDLAND, F. E. THAU, P. E. SARACHIK, "Stability Problems in Randomly-Excited Dynamic Systems", Proc. JACC, Seattle, Wash., 848-861, 1966.

[4] K. ITO, "On Stochastic Differential Equations", Mem. Am. Math. Soc. 4 (1951).

[5] E. B. DYNKIN, "Markov Processes -I, -II", Springer Verlag, Berlin (1965)

[6] S. BERNSTEIN, "Equations Différentielles Stochastiques", Actualités Sci. et Ind. 738, 5-32 (1938).

[7] S. BERNSTEIN, "Correction d'une Démonstration", Comptes Rendus de l' Académie des Sciences de l' URSS, 25, N°9, 707-709 (1939).

[8] G. MARUYAMA, "Continuous Markov Processes and Stochastic Equations", Rendiconti Circolo Mat. Palermo, 4, 1-43 (1955).

[9] I. G. GIHMAN, "On the Theory of Differential Equations of Stochastic Processes", (transl.) Am. Math. Soc. Transl., Ser. 2, 1, 111-161 (1955).

[10] B. V. GNEDENKO, "The Theory of Probability", Chelsea, New York (1962).

[11] L. S. PONTRYAGIN a. o., "The Mathematical Theory of Optimal Processes", Interscience, New York (1962).

[12] R. L. STRATONOVICH, "Topics in the Theory of Random Noise -I", Gordon and Breach, New York (1963).

[13] A. A. SAMARSKII, "On an Economical Method for the Solution of a Multidemensional Parabolic Equation in an Arbitrary Region", (transl.) USSR Computational Mathematics and Mathematical Analysis, 5, 894-926 (1963).

[14] P. DORATO, R. F. DRENICK, L. SHAW, "Optimal Stochastic Control, A Short Course", Polytechnic Institute of Brooklyn, New York (1964).

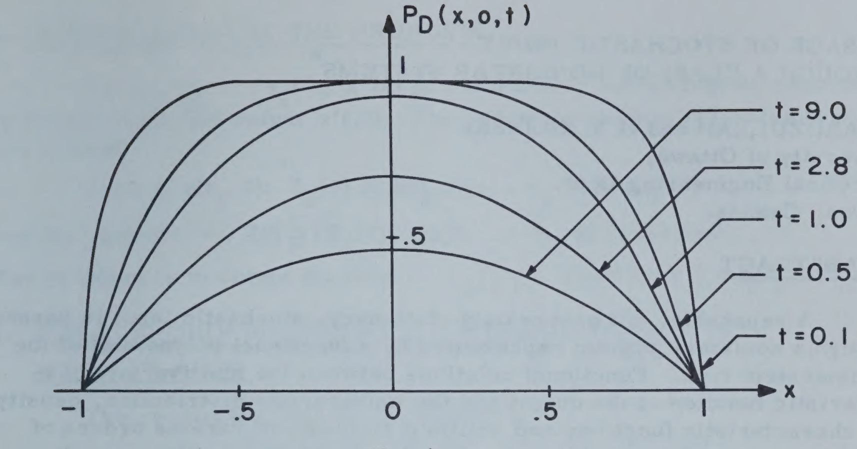

P_D(x,o,t)
1
.5
t=9.0
t=2.8
t=1.0
t=0.5
t=0.1
x
-1
-.5
0
.5
1

Fig. 1

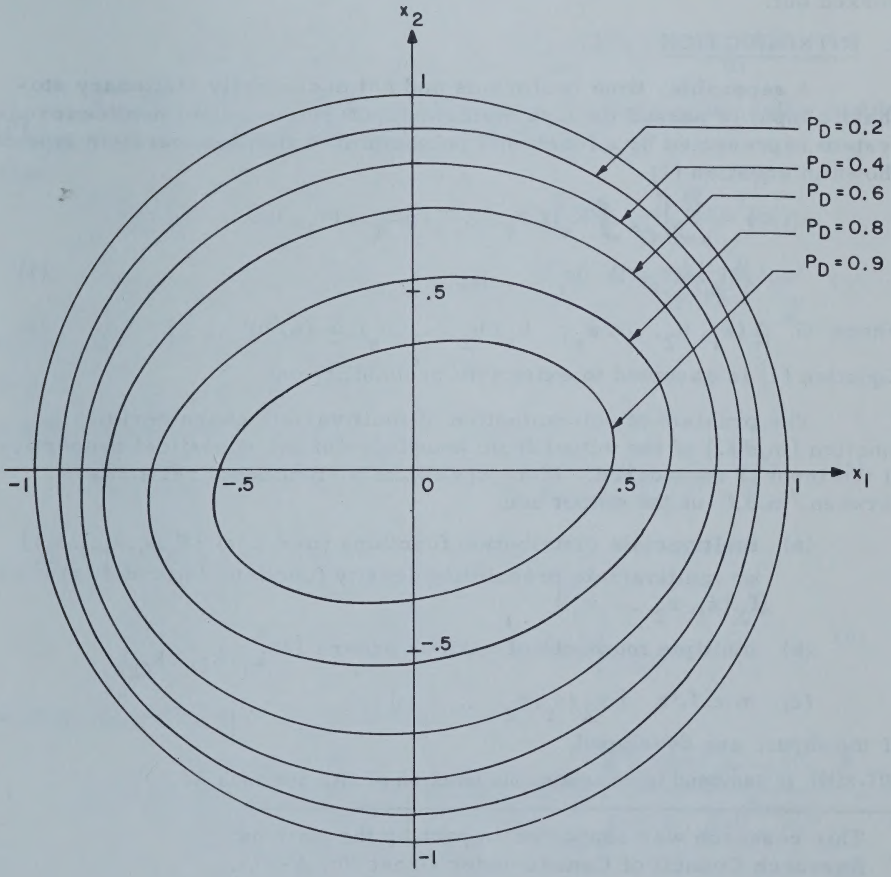

x_2
1
.5
-.5
-1
P_D=0.2
P_D=0.4
P_D=0.6
P_D=0.8
P_D=0.9
x_1
-1
-.5
0
.5
1

Fig. 2

PASSAGE OF STOCHASTIC INPUT THROUGH A CLASS OF NONLINEAR SYSTEMS*.

M. AROZULLAH and G. S. GLINSKI
University of Ottawa,
Electrical Engineering Dept. ,
Ottawa, Canada.

1. ABSTRACT

A separable, not necessarily stationary, stochastic input is passed through a nonlinear system represented by a functional polynomial of the Hammerstein type. Functional relations between the multivariate characteristic function of the output and the multivariate distribution, density and characteristic functions and multiple moments of various orders of the input are developed. Methods of evaluating these relations are discussed. An example of an amplitude modulated communication system is worked out.

2. INTRODUCTION

A separable, time continuous and not necessarily stationary stochastic input is passed through a single-input-single-output nonlinear system represented by a functional polynomial of the Hammerstein type[9] shown in equation (1).

$$z(t,\sigma) = \sum_{s=o}^{N} \int_{G^s} \cdot \int K_s(t,\tau_1, \ldots \tau_s)\, g(\tau_1, x(\tau_1,\sigma)) \ldots \ldots g(\tau_s, x(\tau_s,\sigma))\, d\tau_1 \ldots d\tau_s . \qquad (1)$$

where $G^s \triangleq (a_1, a_2, \ldots . a_s;\ b_1, b_2, \ldots . b_s) \triangleq (a, b)$ (2)

Equation (1) is assumed to exist with probability one.

The problem of determination of multivariate characteristic function (m.c.f.) of the output from knowledge of the statistical properties of the input is considered. More specifically, functional relations between m.c.f. of the output and

(a) multivariate distribution functions (m.d.f.'s) $\{F_X(x_1, x_2, \ldots x_n)\}$ or multivariate probability density functions (m.p. d. f.'s) $\{f_X(x_1, x_2, \ldots x_n)\}$

(b) multiple moments of various orders $\{M^X_{k_1, k_2 \ldots k_Q}\}$

(c) m.c. f.'s $\{\phi_X(\rho_1, \rho_2, \ldots . \rho_n)\}$

of the input, are developed.

$g(t, x(t))$ is assumed to be continuous function of $x(t)$ for each t.

---

* This research was supported in part by the National Research Council of Canada under Grant No. A-875.

3. FORMULATION OF THE PROBLEM

Let $(t_1, t_2, \dots, t_n)$ be a set of values of t satisfying the separability condition for the output $z(t,\sigma)$. The following random variables (r.v.'s) are defined:

$$Z_1(\sigma) \triangleq z(t_1, \sigma);\ Z_2(\sigma) \triangleq z(t_2,\sigma);\ \dots\ Z_n(\sigma) \triangleq z(t_n, \sigma) \tag{3}$$

and the vector r.v., $Z(\sigma) \triangleq (Z_1(\sigma), Z_2(\sigma), \dots Z_n(\sigma))$ is formed.

The problem is to obtain the m.c.f. $\phi_Z(\underline{\xi})$ of $Z(\sigma)$ where $\underline{\xi} = (\xi_1, \xi_2, \dots \xi_n)$.

4. RELATIONSHIP BETWEEN $\phi_Z(\underline{\xi})$ AND $\phi_Y(L,\underline{T})$, m.c.f. of $y(t,\sigma) = g(t, x(t,\sigma))$

The required m.c.f. of the output is

$$\phi_Z(\underline{\xi}) = E\left[e^{i\underline{\xi}^{\dagger} Z(\sigma)}\right] = E\left[\sum_{m=0}^{\infty} \frac{i^m}{m!} \left(\sum_{j_0=1}^{n} \xi_{j_0} Z_{j_0}(\sigma)\right)^m\right]$$

$$= \sum_{m=0}^{\infty} \frac{i^m}{m!} \left(\sum_{j_1=1}^{n} \dots \sum_{j_m=1}^{n} \xi_{j_1} \dots \xi_{j_m} E\left[Z_{j_1}(\sigma) \dots Z_{j_m}(\sigma)\right]\right) \tag{4}$$

where $i^2 = \sqrt{-1}$; $E[\cdot]$ stands for the statistical average of the entity inside the brackets and the dagger $\dagger$ means transposition.

Also,

$$E[Z_{j_1}(\sigma) \dots Z_{j_m}(\sigma)] = \sum_{s_1=0}^{N} \dots \sum_{s_m=0}^{N} \int \dots \int_{G^Q} K_{s_1 \dots s_m}(t_{j_p}, \tau_p^l;$$

$$p = 1, 2, \dots m;\ l = 1, 2 \dots s_p) E\left[\prod_{p=1}^{m} \prod_{j=1}^{s_p} y(\tau_p^j, \sigma)\right] \prod_{p=1}^{m} \prod_{j=1}^{s_p} d\tau_p^j \tag{5}$$

where, $y(\tau_1\ \sigma) \triangleq g(\tau_1, x(\tau_1, \sigma))$, (6)

$Q \triangleq s_1 + s_2 + \dots + s_m$

and the kernels are given by

$$K_{s_1 \dots s_m}(t_{j_p}, \tau_p^l;\ p = 1, 2, \dots m;\ l = 1, 2 \dots s_p) = \prod_{p=1}^{m} K_{s_p}(t_{j_p}, \tau_p^1 \dots \tau_p^{s_p}) \tag{7}$$

Further, for a fixed value of $\underline{T} \triangleq (\tau_1^1, \tau_1^2, \dots \tau_m^{s_m})$,

$$E\left[\prod_{p=1}^{m} \prod_{j=1}^{s_p} y(\tau_p^j, \sigma)\right] = \frac{(i)^{-Q} \partial^Q}{\partial l_1^1 \dots \partial l_m^{s_m}} \phi_y(L, \underline{T}) \Big|_{L=0} \tag{8}$$

with the m.c.f. of the intermediate process,

$$\phi_Y(L,\underline{T}) = \int \dots \int_{G^Q(\underline{T})} \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i l_p^j g(\tau_p^j, x(\tau_p^j))\} \frac{\partial^Q F_X(\underline{T}, x(\tau_1^1), \dots x(\tau_m^{s_m}))}{\partial x(\tau_1^1) \dots \partial x(\tau_m^{s_m})}. \tag{9}$$

where, $F_X(\underline{\tau}, x(\tau_1^1), \ldots, x(\tau_m^{s_m}))$ is the m.d.f. of the input $x(t,\sigma)$ with $\underline{\tau}$ as a parameter and

$$G^Q(\underline{\tau}) \triangleq (a_1(\underline{\tau}), a_2(\underline{\tau}), \ldots, a_Q(\underline{\tau});\ b_1(\underline{\tau}), b_2(\underline{\tau}), \ldots\ b_Q(\underline{\tau}))$$

Putting equation (8) in equation (5) one gets,

$$E[Z_{j_1}(\sigma)\ldots Z_{j_m}(\sigma)] = \sum_{s_1=0}^{N} \cdots \sum_{s_m=0}^{N} \int_{G^Q}\!\cdot\!\int K_{s_1\ldots s_m}(t_{j_p}, \tau_p^l;\ p=1,2,\ldots m;$$

$$l = 1,2,\ldots s_p)\left[\frac{(i)^{-Q}\partial^Q \phi_Y(L,\underline{\tau})}{\partial l_1^1 \ldots \partial l_m^{s_m}}\right]_{L=0} \prod_{p=1}^{m}\prod_{j=1}^{s_p} d\tau_p^j \qquad (10)$$

Using equation (10) in equation (4), one gets,

$$\phi_Z(\underline{\xi}) = \sum_{m=0}^{\infty} \frac{i^m}{m!}\left[\sum_{j_1=1}^{n}\ldots\sum_{j_m=1}^{n} \xi_{j_1}\ldots\xi_{j_m}\left\{\sum_{s_1=0}^{N}\ldots\sum_{s_m=0}^{N}\int\ldots_{G^Q}\int K_{s_1\ldots s_m}(t_{j_p}, \tau_p^l;\right.\right.$$

$$p = 1,2,\ldots m;\ l = 1,2,\ldots s_p)\left(\frac{(i)^{-Q}\partial^Q\phi_Y(L,\underline{\tau})}{\partial l_1^1\ldots\partial l_m^{s_m}}\Bigg|\ L=0\right)$$

$$\left.\left.\prod_{p=1}^{m}\prod_{j=1}^{s_p} d\tau_p^j\right\}\right] \qquad (11)$$

From equation (11) it is seen that the term within the second brackets represents a sum of linear transformations of $\phi_Y(L,\underline{\tau})$, in several variables. Each component linear transformation consists of two parts. The first part gives partial derivative in several variables evaluated at zero values of these variables. The second part is a linear integral transformation. The kernel of this second part is a product of the kernels of the original system represented by equation (1), and is given by equation (7). Different component linear transformations of the sum have kernels which are functions of different number of variables depending on which values $s_1, s_2, \ldots s_m$ take. Thus, equation (11) defines the functional relation between $\phi_Z(\underline{\xi})$ and $\phi_Y(L,\underline{\tau})$ and this relation is linear.

5. DETERMINATION OF $\phi_Y(L,\underline{\tau})$.

It is noted that the Lebesgue-Stieltjes integral in equation (9), defining $\phi_Y(L,\underline{\tau})$, exists. If the input is of a mixed type, then

$$f_X(\underline{\tau}, x(\tau_1^1), \ldots x(\tau_m^{s_m})) = c_1 f_{1_X}(\underline{\tau}, x(\tau_1^1), \ldots x(\tau_m^{s_m}))$$

$$+ c_2 f_{2_X}(\underline{\tau}, x(\tau_1^1), \ldots x(\tau_m^{s_m})) \qquad (12)$$

where $f_X(\underline{\tau}, x(\tau_1^1), \ldots, x(\tau_m^{s_m}))$ is, for each $\underline{\tau}$, the m.p.d.f. of the input, $f_{1_X}(\underline{\tau}, x(\tau_1^1), \ldots, x(\tau_m^{s_m}))$ is the m.p.d.f. corresponding to the discrete portion of $f_X(\underline{\tau}, x(\tau_1^1), \ldots x(\tau_m^{s_m}))$ and is given by

$$f_{1_X}(\underline{T}, x(\tau_1^1), \ldots.\, x(\tau_m^{s_m})) = \sum_{i_1^1=0}^{n_1^1} \ldots. \sum_{i_m^{s_m}=0}^{n_m^{s_m}} p(\underline{T})_{i_1^1 \ldots i_m^{s_m}}$$

$$\prod_{p=1}^{m} \prod_{j=1}^{s_p} \delta(x(\tau_p^j)-(x(\tau_p^j)\, i_p^j), \qquad (13)$$

and $f_{2_X}(\underline{T},\ x(\tau_1^1), \ldots.\, x(\tau_m^{s_m}))$ is the m.p.d.f. corresponding to the continuous portion of the input distribution. Also $c_1, c_2 \geqslant 0$ and $c_1 + c_2 = 1$.

The following cases are considered:

5.1 CASE I - RELATIONSHIP BETWEEN $\phi_Y(L, \underline{T})$ AND $\{F_X(\underline{T}, x(\tau_1^1), \ldots.\, x(\tau_m^{s_m}))\}$ OR $\{f_X(\underline{T}, x(\tau_1^1), \ldots.\, x(\tau_m^{s_m}))\}$.

From equation (9) it follows that

$$\phi_Y(L, \underline{T}) = \int_{G^Q(T)} \ldots \int \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i l_p^j\, g(\tau_p^j, x(\tau_p^j))\} \quad f_X(\underline{T}, x(\tau_1^1), \ldots.\, x(\tau_m^{s_m}))$$

$$\prod_{p=1}^{m} \prod_{j=1}^{s_p} dx(\tau_p^j) \qquad (14)$$

Equation (14) defines the transformation that takes $f_X(\underline{T}, x(\tau_1^1), \ldots x(\tau_m^{s_m}))$ into $\phi_Y(L, \underline{T})$. It is noted that this transformation is linear and its kernel is

$$\prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\ g(\tau_p^j,\ x(\tau_p^j))\}$$

Thus, this kernel is known if g(t, x(t)) is known.

For evaluation of the integral in equation (14), the discrete and continuous portions of the input are considered separately and equation (12) is used in equation (14). Then,

$$\phi_Y(L, \underline{T}) = c_1 \int_{G^Q(\underline{T})} \ldots \int \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\} f_{1_X}(\underline{T}, x(\tau_1^1), \ldots\, x(\tau_m^{s_m}))$$

$$\prod_{p=1}^{m} \prod_{j=1}^{s_p} dx(\tau_p^j) + c_2 \int_{G^Q(\underline{T})} \ldots \int \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\} f_{2_X}(\underline{T}, x(\tau_1^1, \ldots\, x(\tau_m^{s_m}))$$

$$\prod_{p=1}^{m} \prod_{j=1}^{s_p} dx(\tau_p^j) \triangleq I_1 + I_2. \text{ -------------- } \qquad (15)$$

Using equation (13) in equation (15) one gets,

$$I_1 = c_1 \sum_{i_1^1=0}^{n_1^1} \ldots. \sum_{i_m^{s_m}=0}^{n_m^{s_m}} p(\underline{T})_{i_1^1, \ldots\, i_m^{s_m}} \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\, g(\tau_p^j, (x(\tau_p^j))i_p^j)\} \quad (16)$$

where $g(\tau_p^j, (x(\tau_p^j))\, i_p^j)$ is the value of $g(\tau_p^j, x(\tau_p^j))$ at $x(\tau_p^j) = (x(\tau_p^j))i_p^j$ and this is known if g(t, x(t)) is known.

Thus I_1 is completely computable for a given $g(t, x(t))$.

The multiple integral in I_2 becomes product of integrals in one dimension, i.e.

$$I_2 = c_2 \prod_{p=1}^{m} \prod_{j=1}^{s_p} \int_{G(\tau)} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\}\, f_p^j(x(\tau_p^j))\, dx(\tau_p^j) \qquad (17)$$

if the input is white, i.e. if

$$f_{2_X}(\underline{T}, x(\tau_1^1), \ldots, x(\tau_m^{s_m})) = \prod_{p=1}^{m} \prod_{j=1}^{s_p} f_p^j(x(\tau_p^j)). \qquad (18)$$

The one dimensional integrals may, for certain $g(t, x(t))$, be evaluated analytically. But almost always a suitable quadrature formula may be applied. Even if the input is not white, it may be possible to put I_2 in the form of equation (17) by means of a suitable transformation which will depend on the properties of $f_{2_X}(\underline{T}, x(\tau_1^1), \ldots x(\tau_m^{s_m}))$. If I_2 cannot be put in the form of equation (17), then one has to evaluate the multiple integrals in equation (15), analytically, if possible or numerically[3], if necessary.

Thus, if the input is of the mixed type, then $\phi_Y(L, \underline{T})$ is obtained as a sum of two parts, I_1 and I_2, where I_1 is given by equation (16) and I_2 by equation (15) or (17), as the case may be. If the input is purely discrete then $\phi_Y(L, \underline{T})$ is equal to I_1 with $c_1 = 1$ and is obtained from equation (16). Similarly, if the input is purely continuous then $\phi_Y(L, \underline{T})$ is equal to I_2 with $c_2 = 1$ and is obtained from equation (15) or (17), as the case may be.

5.2 CASE II - RELATIONSHIP BETWEEN $\phi_Y(L, \tau)$ AND $\{M^X_{k_1 \ldots k_Q}\}$.

It is noted that $\prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\}$ is a continuous function of $x(\tau_p^j)$, $p = 1, 2, \ldots m$; $j = 1, 2, \ldots s_p$. Use would be made of the Stone-Weierstrass theorem[1, sec 1.1] on the representation of continuous function of real variables. According to this theorem, there exists a sequence of polynomials

$$P_{n_1, \ldots, n_Q}(x(\tau_1^1), \ldots x(\tau_m^{s_m})) = \sum_{k_1=0}^{n_1} \ldots \sum_{k_Q=0}^{n_Q} \gamma_{k_1, \ldots k_Q}(L, \underline{T})(x(\tau_1^1))^{k_1} \ldots \ldots (x(\tau_m^{s_m}))^{k_Q} \qquad (19)$$

such that,

$$\lim_{n_q \to \infty} P_{n_1, \ldots n_Q}(x(\tau_1^1), \ldots x(\tau_m^{s_m})) = \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\} \qquad (20)$$

$q = 1, 2, \ldots Q$

uniformly in $G^Q(\underline{T})$ for almost all L and $\underline{T}$.

Putting equation (20) in equation (14) one gets

$$\phi_Y(L, \underline{T}) = \int \ldots \int_{G^Q(\underline{T})} \lim_{\substack{n_q \to \infty \\ q = 1, 2, \ldots Q}} P_{n_1 \ldots n_Q}(x(\tau_1^1), \ldots x(\tau_m^{s_m}))\, f_X(\underline{T}, x(\tau_1^1), \ldots \ldots x(\tau_m^{s_m})) \prod_{p=1}^{m} \prod_{j=1}^{s_p} dx(\tau_p^j) \qquad (21)$$

The convergence in equation (21) is uniform and hence the order of integration and limiting operation can be interchanged.

Putting equation (19) in equation (21) one gets,

$$\phi_Y(L,\underline{\tau}) = \lim_{n_q \to \infty} \sum_{k_1=0}^{n_1} \cdots \sum_{k_Q=0}^{n_Q} \gamma_{k_1 \ldots k_Q}(L,\underline{\tau})\, M^X_{k_1,\ldots,k_Q} \qquad (22)$$

$$q = 1,2,\ldots Q$$

Equation (22) gives $\phi_Y(L,\underline{\tau})$ as a linear combination of the multiple moments of various orders of the input. Hence, this equation may be used to evaluate $\phi_Y(L,\underline{\tau})$ when the moments are specified. The limiting process $n_q \to \infty$, may not be necessary in most cases and reasonably high n_q's, $q = 1,2,\ldots Q$ are sufficient, depending on the properties of $\prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\}$. The Stone-Weierstrass theorem ensures the existence of the polynomial $P_{n_1,\ldots,n_Q}(x(\tau_1^1),\ldots\, x(\tau_m^{s_m}))$, but one still needs a method of actually determining it. It is noted that $\prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp \{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j))\}$ is the product of functions of single variable. So, some suitable interpolation formula in one dimension, e. g., Lagrangian interpolation formula, Newton's interpolation formula, Aitken-Neville interpolation formula [2, 3], or several other interpolation formulas, may be used to approximate each component function of the product function. Then the required polynomial $P_{n_1 \ldots n_Q}(x(\tau_1^1),\ldots.\, x(\tau_m^{s_m}))$ may be obtained by multiplying the component approximating polynomials.

5.3 CASE III - RELATIONSHIP BETWEEN $\phi_Y(L,\underline{\tau})$ AND $\phi_X(\underline{\rho},\underline{\tau})$.

In this case the relationship between $f_X(\underline{\tau}, X(\tau))$ and $\phi_X(\underline{\rho},\underline{\tau})$ shown in equation (23) is made use of in equation (14)

$$f_X(\underline{\tau}, X(\underline{\tau})) = \int \underset{R^Q}{\cdots} \int \phi_X(\underline{\rho},\underline{\tau})\, e^{-i\underline{\rho}^\dagger X(\underline{\tau})}\, d\underline{\rho} \qquad (23)$$

where $X(\underline{\tau}) \triangleq (x(\tau_1^1),\ldots.\, x(\tau_m^{s_m}))$

Putting equation (23) in equation (14) and interchanging the order of integrations, one gets

$$\phi_Y(L,\underline{\tau}) = \int \underset{R^Q}{\cdots} \int K(L,\underline{\rho},\underline{\tau})\, \phi_X(\underline{\rho},\underline{\tau})\, d\underline{\rho}\,. \qquad (24)$$

where $K(L,\underline{\rho},\underline{\tau}) = \prod_{p=1}^{m} \prod_{j=1}^{s_p} \int_{G(\tau)} \exp\{i\, l_p^j\, g(\tau_p^j, x(\tau_p^j)) - i\, \rho_p^j\, x(\tau_p^j)\}\, dx(\tau_p^j)$ (25)

Equation (24) defines, for each $\underline{\tau}$, a linear transformation of $\phi_X(\underline{\rho},\underline{\tau})$ into $\phi_Y(L,\underline{\tau})$. The kernel $K(L,\underline{\rho},\underline{\tau})$ of this transformation is given by equation (25). From equation (25) it is seen that $K(L,\underline{\rho},\underline{\tau})$ is equal to a product of integrals in one dimension and each integrand is a continuous function of the respective variable of integration. These integrals may, for certain $g(t,x(t))$, be evaluated analytically. But in most cases, they may be evaluated by using certain quadrature formula e.g. Gauss-Legendre quadrature formula.

For evaluation of $\phi_Y(L, \underline{T})$ from equation (24) it is noted that $K(L, \underline{\rho}, \underline{T})$ is a continuous function of $\underline{\rho}$ for almost all $L \in R^{\Omega}$ and all $\underline{T}$ under consideration [4, lemna 2.5] and also is a product of functions of single variable. Hence $\phi_Y(L, \underline{T})$ reduces to product of integrals in single variable if $\phi_X(\underline{\rho}, \underline{T})$ factors into functions of single variable. Then, these integrals in single variable may be evaluated either analytically, if possible, or numerically, if necessary. Otherwise, evaluation of $\phi_Y(L,\underline{T})$ from equation (24) needs evaluation of multiple integrals.

6. AN EXAMPLE

6.1 INTRODUCTION

The amplitude modulated communication link shown in figure 1 is considered. Kac-Siegart [5], Emerson [6], Middleton [7] and others have considered the part of such a link from the I.F. amplifier to the audio amplifier. They assumed the input to the I.F. amplifier to be Gaussian and determined statistical properties of the output of the audio amplifier. In this paper the statistical properties of the output $v(t,\sigma,\alpha,\nu)$ of the square law device are considered. The linear low pass filter is combined with the I.F. amplifier to form a new linear system. Further, it is assumed that the amplifiers preceding and following the channel are not necessary. Thus the system under consideration is as shown in figure 2.

Characteristics of the elementary systems in figure 2 are as follows:

1) The nonlinear system is represented by

$$
\begin{aligned}
y(t, \sigma) &= a_1 e(t, \sigma) + a_2 e^2(t, \sigma). \\
&= a_2 x^2(t, \sigma) + (a_1 + 2a_2 \cos\omega_c t) x(t, \sigma) + a_1 \cos\omega_c t + a_2 \cos^2 \omega_c t \qquad (26)
\end{aligned}
$$

2) system function of the bandpass filter is shown in figure 3, where ω_c is the carrier frequency and ω_b is the maximum frequency component contained in the input.

3) The channel is a simple time-varying channel which can be modelled as in figure 4. This model may represent a channel whose output consists of a directly received ground wave and a sky wave reflected from a moving object, each wave being attenuated differently. Thus the output is [8]

$$
z(t, \sigma) = b_1\, y(t-\beta_1, \sigma) + b_2(\alpha)\, y(t-kt, \sigma) \qquad (27)
$$

where $kt = \beta_2$ is a time varying delay and $b_2(\alpha)$ is a random delay with known p.d.f.

The corresponding system function is

$$
H_c(j\omega, t, \alpha) = b_1 e^{-j\omega\beta_1} + b_2(\alpha) e^{-j\omega kt} \qquad (28)
$$

The composite system function of the combined system of the bandpass filter and the channel is

$$H(j\omega, t, a) = \begin{cases} Ab_1 e^{-j\omega(t_o+\beta_1)} + A\,b_2(a) e^{-j\omega(t_o+kt)} & \text{for } |\omega_c - \omega_b| \le |\omega| \le |\omega_c + \omega_b| \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

The corresponding impulse response is

$$h(t, \tau, a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(j\omega, t, a) e^{j\omega(t-\tau)}\, d\omega$$

$$= \frac{2Ab_1\omega_b}{\pi} \frac{\sin\omega_b(t-\tau-t_o-\beta_1)}{\omega_b(t-\tau-t_o-\beta_1)} \cos\omega_c(t-\tau-t_o-\beta_1)$$

$$+ \frac{2A\,b_2(a)\omega_b}{\pi} \frac{\sin\omega_b(t-\tau-t_o-kt)}{\omega_b(t-\tau-t_o-kt)} \cos\omega_c(t-\tau-t_o-kt). \tag{30}$$

6.2 DETERMINATION OF $\phi_V(\underline{\xi})$, THE m.c.f. OF OUTPUT $v(t, \sigma, a, \nu)$.

The output of the channel is

$$z(t, \sigma, a) = \int_o^T h(t, \tau, a)\, y(t-\tau, \sigma)\, d\tau \tag{31}$$

where T is the memory time of the system.

Then the output of the overall system is

$$v(t, \sigma, a, \nu) = (z(t, \sigma, a) + n(t, \nu))^2$$

$$= \sum_{s=o}^{2} \int_o^T \cdots \int_o^T K_s(t, \tau_1, \ldots, \tau_s, a, \nu) y(t-\tau_1, \sigma) \ldots y(t-\tau_s, \sigma) d\tau_1 \ldots d\tau_s \tag{32}$$

where the kernels of the overall system are

$$\begin{aligned} &(1)\ K_o(t, a, \nu) = n^2(t, \nu) \\ &(2)\ K_1(t, \tau_1, a, \nu) = 2n(t, \nu)\, h(t, \tau_1, a) \\ &(3)\ K_2(t, \tau_1, \tau_2, a, \nu) = h(t, \tau_1, a)\, h(t, \tau_2, a) \end{aligned} \tag{33}$$

Define,

$$V(\sigma, a, \nu) = (V_1(\sigma, a, \nu), \ldots, V_n(\sigma, a, \nu)) \triangleq (V(t_1, \sigma, a, \nu), \ldots V(t_n, \sigma, a, \nu)) \tag{34}$$

Then, the desired m.c.f. of the output of the overall system is

$$\phi_V(\underline{\xi}) = \sum_{m=o}^{\infty} \frac{i^m}{m!} \Big[\sum_{j_1=1}^{n} \cdots \sum_{j_m=1}^{n} \xi_{j_1} \cdots \xi_{j_m} \Big\{ \sum_{s_1=0}^{2} \cdots \sum_{s_m=0}^{2}$$

$$\int_0^T \cdots \int_0^T E\,[K_{s_1 \ldots s_m}(t_{j_p}, \tau_p^l;\ p=1,2,\ldots m;\ l=1,2,\ldots s_p; a, \nu)]$$

$$\Big(\frac{i^{-Q}\,\partial^Q}{\partial l_1^{s_1} \ldots \partial l_m^{s_m}} \phi_Y(L, \underline{T}) \Big|_{L=o} \Big) \prod_{p=1}^{m} \prod_{j=1}^{s_p} d\tau_p^j \Big\}\Big] \tag{35}$$

where

$$(1)\quad K_{s_1 \dots s_m}(t_{j_p}, \tau_p^l;\ p=1,2,\dots m;\ l=1,2,\dots s_p;\ \alpha, \nu) = \prod_{p=1}^{m} K_{s_p}(t_{j_p}, \tau_p^1, \dots \tau_p^{s_p}, \alpha, \nu) \tag{36}$$

with the components of the product given by equation (33).
and

$$(2)\quad E\,[K_{s_1 \dots s_m}(t_{j_p}, \tau_p^l;\ p = 1, 2, \dots m;\ l = 1, 2, \dots s_p;\ \alpha, \nu)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \{K_{s_1 \dots s_m}(t_{j_p}, \tau_p^l; p=1,2,\dots m;\ l=1,2,\dots s_p;\ \alpha, \nu) \Big/ \begin{bmatrix} b_2 \le b_2(\alpha) \le b_2 + db_2 \\ n(t) \le n(t,\nu) \le n(t) + dn(t) \end{bmatrix} \}$$

$$f(b_2)\, f(n_1, \dots, n_k)\, db_2\, dn_1 \dots dn_k\,. \tag{37}$$

where $f(b_2)$ is the p.d.f. of $b_2(\alpha)$ and $f(n_1, \dots n_k)$ is the m.p.d.f. of $n(\nu) = (n_1(\nu), \dots n_k(\nu)) \triangleq (n(t_1, \nu), \dots n(t_k, \nu))$ such that $(t_1, t_2, \dots t_k)$ satisfy the separability condition for the separable stochastic process $n(t, \nu)$. Equation (35) together with equations (36) and (37) gives the required relation between the m.c.f. of the output $v(t, \sigma, \alpha, \nu)$ and that of the intermediate process $y(t, \sigma)$.

6.3 DETERMINATION OF $\phi_Y(L, \underline{T})$, THE m.c.f. OF THE INTERMEDIATE PROCESS $y(t, \sigma)$.

It is assumed that the m.c.f. $\phi_X(\underline{\rho}, \underline{T})$ of the input is given and it is necessary to find $\phi_Y(L, \underline{T})$. Equations (24) and (25) are used here. Equation (24) defines the transformation that takes the m.c.f. $\phi_X(\underline{\rho}, \underline{T})$ of the input into the m.c.f. $\phi_Y(L, \underline{T})$ of the intermediate process. This transformation is defined by the kernel $K(L, \underline{\rho}, \underline{T})$ which is determined next. In this example

$$g(t, x(t)) = a_2 x^2(t) + b(t)x(t) + c(t) \tag{38}$$

$$\text{where } b(t) \triangleq a_1 + 2a_2 \cos \omega_c t \tag{39}$$

$$\text{and } c(t) \triangleq a_1 \cos \omega_c t + a_2 \cos^2 \omega_c t\,.$$

Also τ_p^j in equation (25) is to be replaced by $(t - \tau_p^j)$ for this example, i.e. $\underline{T} = (t - \tau_1^1, \dots t - \tau_m^{s_m})$.

Using equation (25) it can be shown [4, sec, 2.6] that

$$K(L, \underline{\rho}, \underline{T}) = \prod_{p=1}^{m} \prod_{j=1}^{s_p} \exp\,[ia_2 l_p^j \{-(\frac{l_p^j(a_1 + 2a_2 \cos \omega_c (t - \tau_p^j)) - \rho_p^j}{2a_2 l_p^j})^2 + \frac{a_1 \cos \omega_c (t - \tau_p^j) + a_2 \cos^2 \omega_c (t - \tau_p^j)}{a_2}\}](a_2 l_p^j)^{-\frac{1}{2}} \sqrt{\frac{\pi}{2}}\,(1+i). \tag{40}$$

Equation (40) gives the kernel $K(L, \underline{\rho}, \underline{T})$ of the linear transformation of $\phi_X(\underline{\rho}, \underline{T})$ into $\phi_Y(L, \underline{T})$. Thus, this transformation is determined for this example.

7. CONCLUSIONS

The functional relations between several statistical properties of the input and m.c.f. of the output have been obtained. Application of these relations requires evaluation of multiple integrals in many cases. Hence suitable methods of evaluating multiple integrals are essential for application of these relations. The results obtained here may be applied to problem of signal detection, decision problem, design of signals with specific statistical properties, etc.

REFERENCES

1. Timan, A. F., "Theory of Approximation of Functions of a Real Variable", Pergamon Press, New York, U.S.A., 1963.

2. Kopal, Z., "Numerical Analysis", John Wiley and Sons, New York, U.S.A., 1955.

3. Berezin, I. S. and N. P. Zhidhov, "Computing Methods", Vol. I, Pergamon Press, New York, U.S.A., 1965.

4. Arozullah, M., "Passage of Stochastic Input Through a Class of Nonlinear Stochastic Systems", Ph. D. Thesis, Department of Electrical Engineering, Ottawa University, Ottawa, Canada, to be submitted.

5. Kac, M. and A. J. F. Siegert, "On the Theory of Noise in Radio Receivers with Square Law Detectors", Journal of Applied Physics 18 (1947) 383.

6. Emerson, R. C., "First Probability Densities for Receivers with Square Law Detectors", Journal of Applied Physics, 24(1953), 1168.

7. Meyer, M. A. and D. Middleton, "On the Distributions of Signals and Noise After Rectification and Filtering", Journal of Applied Physics, 25 (August, 1954) 1037.

8. Baghdady, E. J. "Lectures on Communication System Theory", McGraw-Hill Book Co., New York, U.S.A., 1961.

9. Ahmed N. U. " Functional Approach To Analysis and Synthesis of Nonlinear Systems", TR 65-14, Electrical Engineering Dept. University of Ottawa, Canada, 1965.

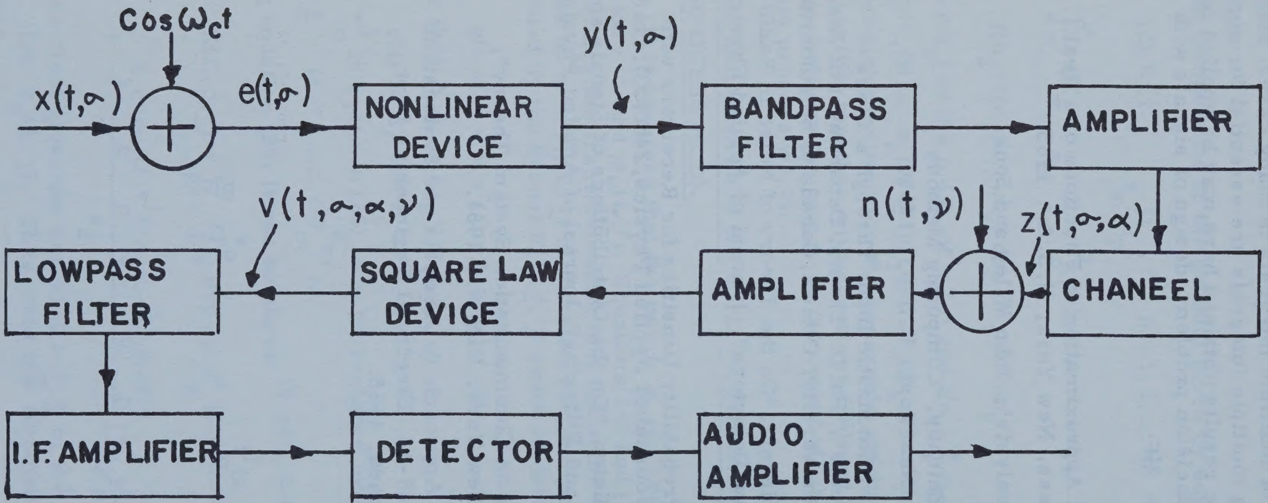

Cos$\omega_c t$
$x(t,\sigma)$
$e(t,\sigma)$
NONLINEAR DEVICE
$y(t,\sigma)$
BANDPASS FILTER
AMPLIFIER
$n(t,\nu)$
$z(t,\sigma,\alpha)$
CHANEEL
AMPLIFIER
SQUARE LAW DEVICE
$v(t,\sigma,\alpha,\nu)$
LOWPASS FILTER
I.F. AMPLIFIER
DETECTOR
AUDIO AMPLIFIER

Fig.1. AMPLITUDE MODULATED COMMUNICATION LINK

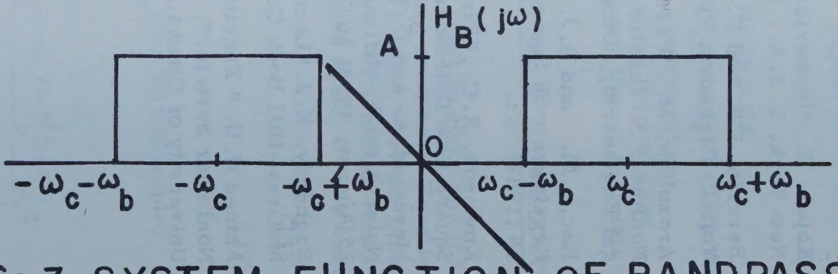


Fig.3. SYSTEM FUNCTION OF BANDPASS FILTER.

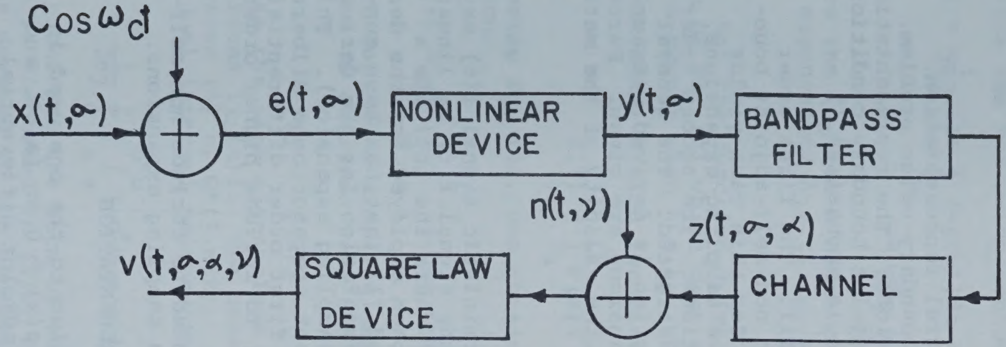

Cos ω_c t
x(t, σ)
e(t, σ)
NONLINEAR DEVICE
y(t, σ)
BANDPASS FILTER
n(t, ν)
z(t, σ, α)
CHANNEL
SQUARE LAW DEVICE
v(t, σ, α, ν)

Fig. 2. MODIFIED AMPLITUDE MODULATED COMMUNICATION LINK.

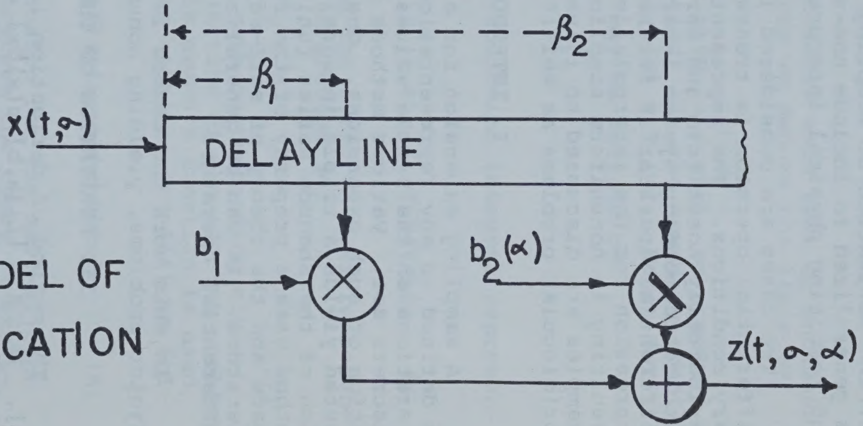

β_2
β_1
x(t, σ)
DELAY LINE
b_1
b_2(α)
z(t, σ, α)

Fig. 4. MODEL OF COMMUNICATION CHANNEL.

INTEGRAL REPRESENTATIONS FOR NONUNIFORM SAMPLING EXPANSIONS

A. H. HADDAD and J. B. THOMAS
Department of Electrical Engineering
Princeton University
Princeton, New Jersey

ABSTRACT

This paper investigates an integral representation, involving the Green's function of a boundary-value problem, for the derivation of sampling expansions. The representation is generalized to include non-self-adjoint boundary conditions, and resulting physical interpretation is emphasized.

Two cases are considered in detail. The first order differential operator is treated with non-self-adjoint boundary conditions. The representation for this case yields periodically nonuniform and derivative sampling expansions, with corresponding system interpretations. The second order differential operator is further investigated. The general expression for the interpolating function is derived. Examples resulting in nonuniform sampling expansions are given. Further examples are discussed to illustrate the validity of the method for singular problems as well.

I. INTRODUCTION

A sampling expansion for a deterministic signal f(t) may be defined as any representation of the signal f(t) by linear operations on the sample values $f(t_i)$ where the $\{t_i\}$ are a discrete set. Various methods have been employed for the derivation of such expansions. One method [1] that has been suggested yields a triple integral representation for the derivation of the Shannon Series (uniform sampling expansion). This method uses a property of the resolvent operator on a Hilbert space and the theory of self-adjoint first order differential operators. It can be generalized [2] to include higher order differential operators.

In this work the method is extended to include non-self-adjoint problems, yielding nonuniform sampling expansions.

II. DERIVATION OF THE REPRESENTATION

The method of derivation is similar to the one used in [2]. Let $H = \{L_2[a,b];\rho(x)dx\}$ where $\rho(x) > 0$ on (a,b), and let L be an n-th order formally self-adjoint differential operator [3,4] with domain $D = \{y \in C^n[a,b]$ and satisfying n independent boundary conditions $Uy = 0\}$. Consider the case where the boundary-value problem

$$Ly - \lambda\rho(x)y = 0 \quad , \quad y \in D \tag{1}$$

has only real eigenvalues $\{\lambda_i\}$, with corresponding eigenfunctions $\{\psi_i(x)\}$. The nonhomogeneous differential equation

$$Ly - z\rho(x)y = \rho(x)F(x) \quad , \quad F(x) \in H \tag{2}$$

has a solution $y \in D$ for $z \neq \lambda_i$ which may be expressed by the

Green's function of the equation as [3]

$$y(x) = \int_a^b G(x,\xi;z)F(\xi)\rho(\xi)d\xi \tag{3}$$

On using a property of the resolvent operator [1,5], one obtains the following contour integral [6]

$$F(x) = \frac{1}{2\pi j} \lim_{m\to\infty} \oint_{C_m} [-y(x)]dz =$$
$$= \frac{1}{2\pi j} \oint_C \int_a^b [-G(x,\xi;z)]F(\xi)\rho(\xi)d\xi dz \tag{4}$$

where the $\{C_m\}$ is a sequence of contours such that C_m encloses the eigenvalues $\lambda_1,\ldots,\lambda_m$, and where we denote the limit by $\oint\limits_C$.

Let the function $\psi_\lambda(x) = \psi(x,t)\in H$, $\lambda = g(t)$, be a solution of Eq. (1) which does not satisfy all the boundary conditions except when $\lambda = \lambda_i$, whence $\psi_{\lambda_i}(x) = \psi_i(x)$. Hence for each $F(x)\in H$, a time function $f(t)$ may be defined by

$$f(t) = \int_a^b F(x)\psi^*(x,t)\rho(x)dx \tag{5}$$

On using Eq. (4), one obtains the following integral representation

$$F(t) = \frac{1}{2\pi j} \oint_C \int_a^b \int_a^b [-G(x,\xi;z)]F(\xi)\psi^*(x,t)\rho(\xi)\rho(x)d\xi dx dz \tag{6}$$

which defines a sampling expansion for $f(t)$. A dual representation is obtained from the adjoint problem [6]

$$Ly - \lambda\rho(x)y = 0 \quad , \quad U^+y = 0 \tag{7}$$

It has the same eigenvalues $\{\lambda_i\}$, with corresponding eigenfunctions $\{\chi_i(x)\}$. The Green's function of the adjoint problem is given [6] by $G^*(\xi,x;z)$. If this Green's function is used, the dual integral representation is obtained for $f(t)$ as follows

$$f(t) = \frac{1}{2\pi j} \oint_C \int_a^b \int_a^b [-G^*(\xi,x;z)]F(\xi)\rho(\xi)\psi^*(x,t)\rho(x)d\xi dx dz \tag{8}$$

The proof and the form of the expansion for the self-adjoint case is discussed in [2]. To illustrate the types of expansions derived in the non-self-adjoint case, assume that the Green's function has only simple poles at $z = \lambda_i$. If each eigenvalue is simple, the residues at these poles are given [6] by $\psi_i(x)\chi_i^*(\xi)$. In this case the representation of Eq. (8) results in the expansion

$$f(t) = \sum_{i=1}^{\infty} f(t_i)\varphi_i(t) \quad , \quad \lambda_i = g(t_i) \tag{9}$$

$$\varphi_i(t) = \int_a^b \chi_i(x)\,\psi^*(x,t)\,\rho(x)\,dx$$

Similarly Eq. (6) yields the expansion

$$f(t) = \sum_{i=1}^{\infty} f_1(t_i)\,\theta_i(t) \quad , \qquad \lambda_i = g(t_i)$$

$$\theta_i(t) = \int_a^b \psi_i(x)\,\psi^*(x,t)\,\rho(x)\,dx \tag{10}$$

The function $f_1(t)$ is obtained from $f(t)$ by a linear time-varying operation, and $\varphi_i(t)$ is obtained from $\theta_i(t)$ by the same operator. The case where the Green's function has poles of multiplicity m_i at $z = \lambda_i$ yields similarly a representation of the form

$$f(t) = \sum_{i=1}^{\infty}\sum_{n=1}^{m_i} f_n(t_i)\,\varphi_{i,n}(t) \tag{11}$$

where the $\{f_n(t)\}$ are derived from $f(t)$ by linear operations.

The novelty of this representation is in the physical interpretations it offers. This interpretation is treated in [1] for the first order self-adjoint case. The systems interpretations for the non-self-adjoint case are illustrated by the particular cases.

III. THE FIRST ORDER DIFFERENTIAL OPERATOR

Let $H = L_2[-\pi,\pi]$ and consider the normalized first order differential operator [7]

$$Ly = j\,\frac{dy}{dx} \tag{12}$$

with domain

$$D = \left\{ y \in C^1[-\pi,\pi],\ \text{and satisfying } \sum_{k=0}^{m} a_k y\!\left(-\pi + k\,\frac{2\pi}{m}\right) = 0 \right\} \tag{13}$$

The eigenvalues of the boundary-value problem

$$j\,\frac{dy}{dx} - \lambda y = 0 \quad , \qquad y \in D \tag{14}$$

are the roots of the trigonometric polynomial $\omega(\lambda)$ given by

$$\omega(\lambda) = \sum_{k=0}^{m} a_k e^{-j\frac{\lambda\pi}{m}(2k-m)} = 0 \tag{15}$$

The constants a_k are to be chosen such that all the zeros of this polynomial are real. The eigenvalues are periodically spaced, and if we denote the first m eigenvalues by $0 \leq \lambda_i < m$, $i = 1,2,\ldots,m$, all the eigenvalues of Eq. (14)

will be given by

$$\lambda_{n,i} = \lambda_i + nm \quad , \quad n = 0, \pm 1, \pm 2, \ldots \tag{16}$$

The Green's function of the nonhomogeneous equation is given by [5]

$$G(x,\xi;z) = \frac{je^{jz(\xi-x)}}{\omega(z)} \cdot \begin{cases} \left[-\sum_{k=0}^{\ell-1} a_k e^{-jz\frac{\pi}{m}(2k-m)}\right], & x > \xi \\ \left[\sum_{k=\ell}^{m} a_k e^{-jz\frac{\pi}{m}(2k-m)}\right], & x \le \xi \end{cases} \tag{17}$$

for $[-\pi+(\ell-1)\frac{2\pi}{m}] \le \xi < [-\pi+\ell\,\frac{2\pi}{m}]$. If we let $\lambda = t$ and use the solution

$$\psi(x,t) = \frac{1}{2\pi}\, e^{-jxt} \tag{18}$$

then Eq. (5) defines a function f(t) band-limited to $(-\pi,\pi)$. The integral representations of Eqs. (6) and (8) will therefore yield a periodically nonuniform sampling expansion for f(t).

Let us consider the case where all the λ_i are distinct and perform the contour integral first. The result will be given as the sum of the residues $R_{n,i}(x,\xi)$ of $[-G(x,\xi;z)]$ at the poles $z = \lambda_{n,i}$. These are given by

$$R_{n,i}(x,\xi) = -\frac{m}{2\pi}\,\frac{\sum_{k=0}^{\ell-1} a_k e^{-j\lambda_i \frac{\pi}{m}(2k-m)}}{\sum_{k=0}^{m} k a_k e^{-j\lambda_i \frac{\pi}{m}(2k-m)}}\, e^{j(\lambda_i+mn)(\xi-x)} =$$

$$= \frac{1}{2\pi} H_i(\xi) e^{j(\lambda_i+mn)(\xi-x)} \quad , \quad [-\pi+(\ell-1)\,\frac{2\pi}{m}] \le \xi < [-\pi+\ell\,\frac{2\pi}{m}] \tag{19}$$

On using this relationship in the representation of Eq. (8) we obtain

$$f(t) = \sum_{n=-\infty}^{\infty} \sum_{i=1}^{m} f(\lambda_i+mn) h_i[t-(\lambda_i+mn)] =$$

$$= \sum_{n=-\infty}^{\infty} \sum_{i=1}^{m} Bf(\lambda_i+mn) h_i[t-(\lambda_i+mn)] \tag{20}$$

where $h_i(\tau)$ is the impulse response of the system whose complex transfer function is $H_i^*(x)$ and where we have introduced the ideal low-pass filter operation on f(t) as Bf(t). The integral representation of Eq. (6) results in the following equation where sinc $\tau = (\sin \pi\tau)/\pi\tau$ is the impulse response of the operator B, and $H_i f(t)$ is the output of the system $h_i(\tau)$ if the input is f(t):

$$f(t) = \sum_{n=-\infty}^{\infty}\sum_{i=1}^{m} H_i f(\lambda_i+mn)\,\mathrm{sinc}[t-(\lambda_i+mn)] \tag{21}$$

The representations obtained in Eqs. (20) and (21) may be illustrated physically as in Fig. 1.

If two or more of the λ_i are equal, derivative sampling expansions may be obtained by this method. In particular, consider the case where all the λ_i in one period are equal. This may be achieved if the constants a_k in the boundary condition are chosen to be

$$a_k = (-1)^k \binom{m}{k} = (-1)^k \frac{m!}{k!(m-k)!} \tag{22}$$

In this case $\omega(\lambda)$ in Eq. (15) becomes

$$\omega(\lambda) = \sum_{k=0}^{m} (-1)^k \binom{m}{k} e^{-j\frac{\lambda\pi}{m}(2k-m)} = \left(2j \sin \frac{\lambda\pi}{m}\right)^m \tag{23}$$

The Green's function of the equation is given by Eq. (17), where the a_k and $\omega(\lambda)$ are given by Eqs. (22) and (23) respectively; it reduces to Eq. (23) in [1] for m = 1. The Green's function has an m-th order pole at $z = \lambda_n = mn$, $n=0,\pm1,\pm2,\ldots$. The residue $R_n(x,\xi)$ at the pole $z = mn$ is given by

$$R_n(x,\xi) = \sum_{i=0}^{m-1} (-jx)^i H_i(\xi) e^{jmn(\xi-x)} \tag{24}$$

where

$$H_{m-1}(\xi) = j\left(\frac{m}{2\pi j}\right)^m \sum_{k=0}^{\ell-1} \frac{(-1)^k m}{k!(m-k)!}\ , \quad -\pi+(\ell-1)\frac{2\pi}{m} \leq \xi < -\pi+\ell\,\frac{2\pi}{m} \tag{25}$$

and the $H_i(\xi)$ satisfy the relationship

$$\frac{dH_i(\xi)}{d(j\xi)} = (i+1)H_{i+1}(\xi)\ , \quad i = 0,1,\ldots,m-2 \tag{26}$$

Define $h_i(\tau)$ as the impulse response corresponding to the system H_i with frequency response $2\pi H_i^*(\xi)$, namely

$$h_i(\tau) = \int_{-\pi}^{\pi} H_i^*(\xi) e^{j\xi\tau} d\xi = \int_{-\pi}^{\pi} H_i(\xi) e^{-j\xi\tau} d\xi \tag{27}$$

In particular, Eq. (25) yields

$$h_{m-1}(\tau) = \frac{\tau^{m-1}}{(m-1)!} \left(\mathrm{sinc}\,\frac{\tau}{m}\right)^m \tag{28}$$

and Eqs. (26), (27), and (28) result in the expression

$$h_i(\tau) = \frac{(i+1)}{\tau} h_{i+1}(\tau) = \frac{\tau^i}{i!}\left(\mathrm{sinc}\,\frac{\tau}{m}\right)^m\ , \quad i = 0,1,\ldots,m-1 \tag{29}$$

To obtain the sampling expansions from the representations in Eqs. (6) and (8), we perform the contour integration first. The representation of Eq. (8) yields the expansion

$$f(t) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{m-1} f^{(i)}(mn) h_i(t-mn) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{m-1} G_i f(mn) h_i(t-mn) \tag{30}$$

where we have defined the systems G_i with impulse responses $g_i(\tau)$ given by the expressions

$$g_i(\tau) = \frac{d^i}{d\tau^i} \operatorname{sinc}(\tau) = \frac{d^i}{d\tau^i} \left(\frac{\sin \pi\tau}{\pi\tau}\right), \quad i = 0,1,\ldots,m-1 \tag{31}$$

Similarly the expansion of the representation in Eq. (6) becomes

$$f(t) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{m-1} H_i f(mn) g_i(t-mn) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{m-1} H_i f(mn) g_0^{(i)}(t-mn) \tag{32}$$

where

$$H_i f(t) = \int_{-\pi}^{\pi} F(x)(-1)^i H_i(x) e^{jxt} dx = \int_{-\infty}^{\infty} f(t-\tau) h_i(\tau) d\tau \tag{33}$$

In this case system representations may be derived from the expansions of Eqs. (30) and (32) and are shown in Fig. 2.

Different system representations may be derived by using different permutations of the integrals in the same manner as for the case m = 1 which is discussed by Yao and Thomas [1].

IV. THE SECOND ORDER DIFFERENTIAL OPERATOR

The case where L is a second order self-adjoint differential operator will now be treated. The operator is given by [8]

$$Ly = q(x)y(x) - \frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] \tag{34}$$

where $y \in C^2[a,b]$ and satisfies the boundary conditions

$$\begin{aligned} y(a)\cos\alpha - y'(a)p(a)\sin\alpha &= 0 \\ y(b)\cos\beta - y'(b)p(b)\sin\beta &= 0 \end{aligned} \tag{35}$$

The choice of the transformation $\lambda = t^2$ results in the sampling expansion discussed by Kramer [9] and Campbell [10]. The representation of Eq. (6) will yield the expansion

$$f(t) = \sum_{i=1}^{\infty} f(t_i)\varphi_i(t)$$
$$\varphi_i(t) = \frac{1}{||\psi_i||^2} \int_a^b \psi_i(x)\psi^*(x,t)\rho(x)dx \tag{36}$$

Here $G(x,\xi;z)$ is the Green's function of the equation

$$-\left\{\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + [z\rho(x) - q(x)]y(x)\right\} = F(x) \tag{37}$$

with y satisfying the boundary conditions of Eq. (35). The function $\psi(x,t)$ is chosen as a solution of the homogeneous equation satisfying only one of the boundary conditions of Eq. (35).

The properties of the eigenvalues λ_i and the functions $\varphi_i(t)$ are sufficient to determine the nature of the expansion of Eq. (36). The self-adjoint case cannot yield derivative sampling since the eigenvalues are distinct. The sampling expansion may be nonuniform, but it is asymptotically uniform. To derive the functions $\varphi_i(t)$ from Eq. (36), we may assume that $\psi(x,t)$ satisfies the first boundary condition of Eq. (35). We obtain

$$||\psi_i||^2 \varphi_i(t) = \int_a^b \psi_i(x)\psi(x,t)\rho(x)\,dx = \tag{38}$$

$$= \frac{1}{\lambda_i} p(b)\,[\psi'(b,t)\psi_i(b) - \psi(b,t)\psi_i'(b)] + \frac{\lambda}{\lambda_i} ||\psi_i||^2 \varphi_i(t)$$

This relationship in turn yields the expression for $\varphi_i(t)$

$$\varphi_i(t) = \frac{p(b)}{||\psi_i||^2(\lambda-\lambda_i)}\,[\psi(b,t)\psi_i'(b) - \psi_i(b)\psi'(b,t)] \tag{39}$$

The numerator of this last expression has as zeros the eigenvalues $\lambda_n = t_n^2$. Furthermore [11] the numerator is an entire function in λ of order at most 1/2. The zeros λ_n of this entire function satisfy the condition [12]

$$\sum_{\lambda_n \neq 0} \lambda_n^{-\frac{1}{2}-\varepsilon} < \infty \quad , \quad \text{all } \varepsilon > 0 \tag{40}$$

Hence $\{\lambda_n\}$ is of genus zero, and, by the Hadamard factorization theorem [13], the numerator may be written as

$$\lambda^m e^{Q(\lambda)} \prod_{n=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_n}\right) \tag{41}$$

where m is the multiplicity of the zero at the origin. In this case the multiplicity m is either zero or unity. The $Q(\lambda)$ is a polynomial of degree less than 1/2, i.e. is a constant. Assuming m = 0, we obtain the expression for $\varphi_i(t)$

$$\varphi_i(t) = \frac{A_i}{\lambda-\lambda_i} \prod_{n=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_n}\right) = \frac{A_i}{t^2-t_i^2} \prod_{n=1}^{\infty} \left(1 - \frac{t^2}{t_n^2}\right) \tag{42}$$

The constant A_i is to be chosen to satisfy the condition $\varphi_i(t_i) = 1$ which results in the final expression

$$\varphi_i(t) = \frac{2t_i}{G'(t_i)} \frac{G(t)}{(t^2-t_i^2)} = \frac{G(t)}{G'(t_i)(t-t_i)} + \frac{G(t)}{G'(-t_i)(t+t_i)} \tag{43}$$

where we have defined

$$G(t) = \prod_{n=1}^{\infty} \left(1 - \frac{t^2}{t_n^2}\right) \tag{44}$$

This result may be summarized by the proposition: The sampling expansion defined by Eq. (6) is equivalent to the Lagrange interpolation formula. Two examples will now be considered.

Example 1 - Expansions in Trigonometric Functions - The special case $q(x) = 0$, $\rho(x) = p(x) = 1$, $[a,b] = [0,\pi]$ will be considered. The differential equation becomes

$$y''(x)+\lambda y(x) = 0 \quad , \quad \lambda = t^2 \tag{45}$$

Three different boundary conditions will be discussed: the first two are special cases of Eq. (35) and the third is a non-self-adjoint boundary condition.

Case I: $\alpha = 0$ - In this case

$$\psi(x,t) = \frac{1}{\pi} \sin xt \tag{46}$$

and the Green's function of the equation is [14]

$$G(x,\xi;z) = \begin{cases} \dfrac{\sin\xi\sqrt{z}}{\sqrt{z}} \; \dfrac{\cos\beta\sin(\pi-x)\sqrt{z}-\sqrt{z}\sin\beta\cos(\pi-x)\sqrt{z}}{\cos\beta\sin\pi\sqrt{z}-\sqrt{z}\sin\beta\cos\pi\sqrt{z}}, & 0\le\xi\le x\le\pi \\ \dfrac{\sin x\sqrt{z}}{\sqrt{z}} \; \dfrac{\cos\beta\sin(\pi-\xi)\sqrt{z}-\sqrt{z}\sin\beta\cos(\pi-\xi)\sqrt{z}}{\cos\beta\sin\pi\sqrt{z}-\sqrt{z}\sin\beta\cos\pi\sqrt{z}}, & 0\le x\le\xi\le\pi \end{cases} \tag{47}$$

The sampling instants $t_i = \sqrt{\lambda_i}$ are given by the poles of the Green's function, and are to satisfy

$$(\tan \pi t_i)/\pi t_i = (\tan \beta)/\pi \tag{48}$$

The representation of Eq. (6) results in a nonuniform sampling expansion for a band-limited odd function given by

$$f(t) = \sum_{i=-\infty}^{\infty} h(t_i) f(t_i) \frac{\sin \pi(t-t_i)}{\pi(t-t_i)} \; , \; t_{-i} = -t_i \tag{49}$$

where

$$h(t) = \frac{1+t^2\tan^2\beta}{1+t^2\tan^2\beta - \dfrac{\tan\beta}{\pi}} \tag{50}$$

Case II: $\alpha = \pi/2$ - The expansion will be defined for a band-limited even function since

$$\psi(x,t) = \frac{1}{\pi} \cos xt \tag{51}$$

The Green's function of the equation is given by a similar expression to Eq. (47), and the sampling instants t_i satisfy the equation

$$(\cot \pi t_i)/\pi t_i = -(\tan \beta)/\pi \tag{52}$$

The resulting expansion is identical to the one in Eq. (49).

Case III: Non-self-adjoint system - The boundary conditions will be:

$$y(0) = 0 \; , \; y'(0)+y'(\pi) = 0 \tag{53}$$

The solution $\psi(x,t)$ is given in Eq. (46). The eigenvalues are $\lambda_i(2i+1)^2$, $i = 0,1,\ldots$ and the Green's function is given by [15]

$$G(x,\xi;z) = \frac{\sin\sqrt{z}x\cos\sqrt{z}(\pi-\xi)}{2\sqrt{z}\cos^2\frac{\pi\sqrt{z}}{2}} - \begin{cases} \dfrac{\sin(x-\xi)\sqrt{z}}{\sqrt{z}}, & 0\le\xi\le x\le\pi \\ 0, & 0\le x\le\xi\le\pi \end{cases} \tag{54}$$

The Green's function has double poles at all the eigenvalues. Therefore the substitution of Eqs. (46) and (54) in the representations (6) and (8) will result in a uniform first-derivative sampling expansion for an odd function and will be given by Eqs. (32) and (30) respectively for $m = 2$. Similar expansions may be obtained for an even function.

Example 2 - Expansion in Gegenbauer Polynomials - The D.E. is

$$\frac{d}{dx}\left\{(1-x^2)^{\nu+\frac{1}{2}}y'(x)\right\}+(t^2-\nu^2)(1-x^2)^{\nu-\frac{1}{2}}y(x) = 0;\ \nu > 0 \tag{55}$$

where

$$\rho(x) = (1-x^2)^{\nu-\frac{1}{2}},\quad [a,b] = [-1,1]$$

Its solution is the Gegenbauer function [16]

$$\psi(x,t) = C^{\nu}_{t-\nu}(x) \tag{56}$$

This D.E. is singular at both ends of the interval [-1,1]; therefore the boundary conditions will be to require the functions $\psi_n(x)$ to be bounded at ±1. A complete set of orthogonal functions is the set of Gegenbauer polynomials [16,17]

$$\psi_n(x) = C^{\nu}_n(x) = \frac{(-2)^n\Gamma(n+\nu)\Gamma(n+2\nu)}{n!\Gamma(\nu)\Gamma(2n+2\nu)}(1-x^2)^{\frac{1}{2}-\nu}\frac{d^n}{dx^n}[(1-x^2)^{n+\nu-\frac{1}{2}}] \tag{57}$$

corresponding to the eigenvalues $\lambda_n=t_n^2=(n+\nu)^2$, $n=0,1,2,\ldots$.

To derive the expansion, we may use Eqs. (36), (56) and (57) directly and obtain finally [5]

$$f(t) = \sum_{n=0}^{\infty} f(n+\nu)\left[\frac{\sin\pi(t-n-\nu)}{\pi(t-n-\nu)} - \frac{\sin\pi(t-n-\nu)}{\pi(t+n+\nu)}\right],\ t > 0 \tag{58}$$

The class of functions $f(t)$ has the form

$$f(t) = \int_{-1}^{1}F(x)C^{\nu}_{t-\nu}(x)(1-x^2)^{\nu-\frac{1}{2}}dx = \pi\frac{\Gamma(t+\nu)}{\Gamma(t-\nu+1)}g(t) \tag{59}$$

where $g(t)$ is an even function band-limited to $(-\pi,\pi)$. The expansion for $g(t)$ will then hold for all t and will be given by

$$g(t) = \sum_{n=0}^{\infty} g(n+\nu)\frac{(-1)^{n+1}\Gamma(n+2\nu)}{n!\Gamma(\nu-t)\Gamma(\nu+t)}\left[\frac{1}{t-n-\nu} - \frac{1}{t+n+\nu}\right] \tag{60}$$

The result is a uniform expansion on the intervals $(-\infty,-\nu]$ and $[\nu,\infty)$, with a gap of width 2ν between the last sample of the left interval and the first sample of the right interval.

Two special values of ν are of interest. The first case is when $\nu = m + 1/2$ where m is an integer or zero. Then the expansion for an even function whose center 2m samples vanish is obtained

$$f(t) = \sum_{n=m}^{\infty} [f(n+\tfrac{1}{2})\,\mathrm{sinc}(t-n-\tfrac{1}{2})+f(-n-\tfrac{1}{2})\,\mathrm{sinc}(t+n+\tfrac{1}{2})] \qquad (62)$$

The second case is when $\nu = m$ (m an integer), which results in the expansion for an odd function whose center (2m+1) samples vanish,

$$f(t) = \sum_{n=m}^{\infty} [f(n)\,\mathrm{sinc}(t-n)+f(-n)\,\mathrm{sinc}(t+n)] \qquad (63)$$

The case $\nu = 1/2$ reduces to the Legendre polynomials and was discussed by Campbell [10].

ACKNOWLEDGMENT

The authors wish to express their appreciation for the generous support of the National Science Foundation under Grants GP-1647 and GK-187, and of the U.S. Army Research Office - Durham, under Contract DA-31-124-ARO-D-292.

REFERENCES

1. K. Yao and J.B. Thomas, "On Some Representations and Sampling Expansions for Band-Limited Signals", Tech. Rept. No. 9, Comm. Lab., Dept. of Elec. Engr., Princeton University, pp. 35-54; April 1965.
2. A.H. Haddad, K. Yao, and J.B. Thomas, "General Methods for the Derivation of Sampling Expansions", to be published in IEEE Trans. on Information Theory.
3. E.A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955; Chap. 7, pp. 186-207.
4. B. Friedman, Principles and Techniques of Applied Mathematics, Wiley, New York, 1964; p. 148.
5. A.H. Haddad and J.B. Thomas, "Integral Representations for the Derivation of Sampling Expansions", Tech. Rept. No. 17, Information Sciences and Systems Lab., Dept. of Elec. Engr., Princeton University, August 1966.
6. E.A. Coddington and N. Levinson, ibid., Chap. 12, pp. 298-313.
7. N.I. Akhiezer and I.M. Glazman, Theory of Linear Operators in Hilbert Space, Ungar, New York, 1961; Vol. 1, pp. 106-113.
8. E.C. Titchmarsh, Eigenfunction Expansions, Oxford University Press, London, 1962; Pt. 1, Chap. 1-2, pp. 1-49.
9. H.P. Kramer, "A Generalized Sampling Theorem", J. Math. and Physics, Vol. 38, pp. 68-72, April 1959.
10. L.J. Campbell, "A Comparison of the Sampling Theorems of Kramer and Whittaker", J. Soc. Indust. App. Math., Vol. 12, pp. 117-130, March 1964.

11. F.V. Atkinson, Discrete and Continuous Boundary Problems, Academic Press, New York, 1964; p. 206.
12. F.V. Atkinson, ibid., p. 208.
13. R.P. Boas, Entire Functions, Academic Press, New York, 1954; pp. 18-22.
14. E.C. Titchmarsh, ibid., pp. 7-8, 16.
15. E.A. Coddington and N. Levinson, ibid., p. 312.
16. A. Erdelyi, W. Magnus, F. Oberhettinger, and F.G. Tricomi, Higher Transcendental Functions, Vol. 1, McGraw-Hill, New York, 1953; pp. 175-179.
17. A. Erdelyi et al., ibid., Vol. 2, pp. 174-178.

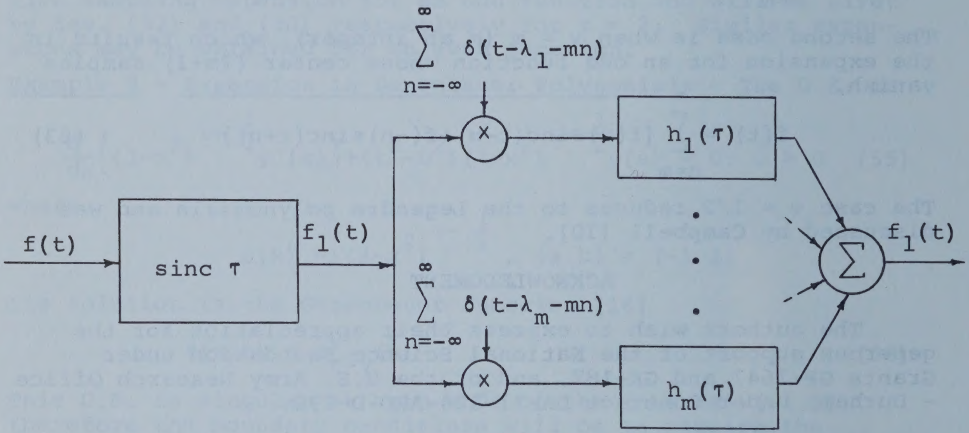

$\sum_{n=-\infty}^{\infty} \delta(t-\lambda_1-mn)$
f(t)
sinc τ
$f_1(t)$
×
$h_1(\tau)$
$\sum_{n=-\infty}^{\infty} \delta(t-\lambda_m-mn)$
×
$h_m(\tau)$
Σ
$f_1(t)$

Fig. 1(a) System Representation for Eq. (20)

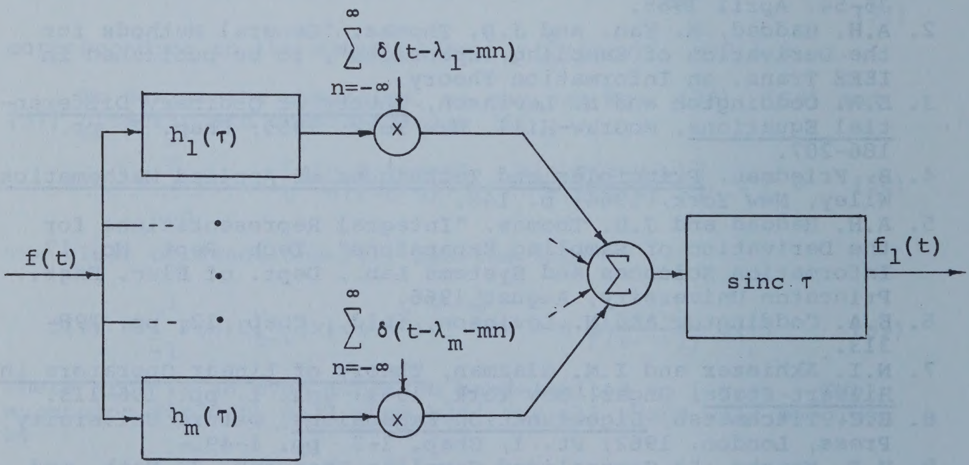
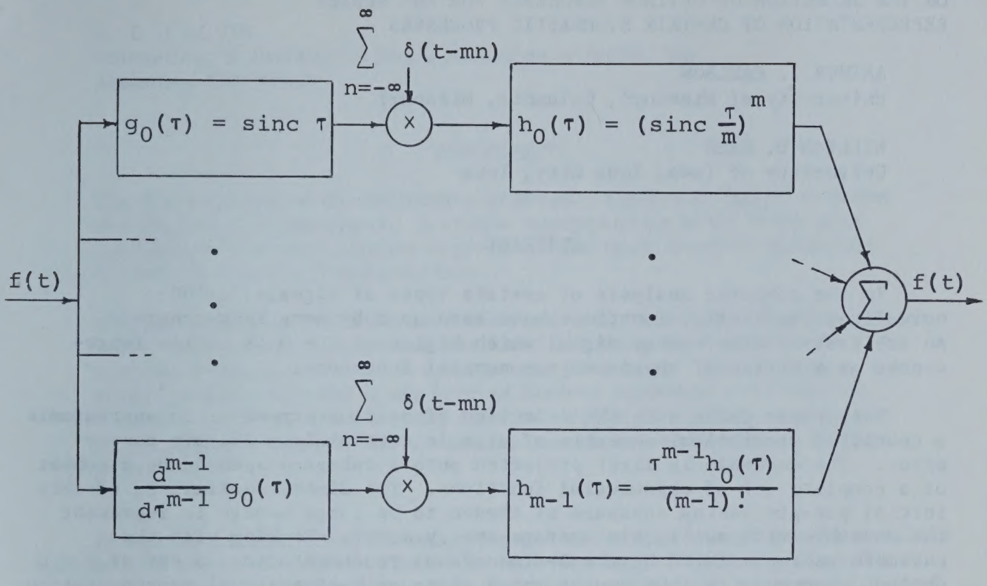
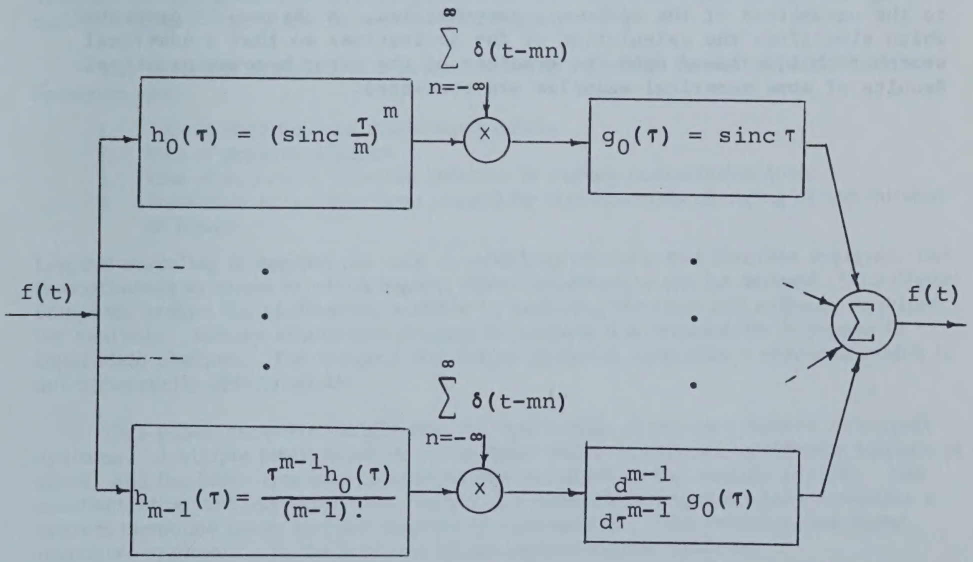

$\sum_{n=-\infty}^{\infty} \delta(t-\lambda_1-mn)$
f(t)
$h_1(\tau)$
×
$\sum_{n=-\infty}^{\infty} \delta(t-\lambda_m-mn)$
$h_m(\tau)$
×
Σ
sinc τ
$f_1(t)$

Fig. 1(b) System Representation for Eq. (21)



f(t)
$g_0(\tau) = \mathrm{sinc}\,\tau$
$\sum_{n=-\infty}^{\infty} \delta(t-mn)$
×
$h_0(\tau) = (\mathrm{sinc}\,\frac{\tau}{m})^m$
$\frac{d^{m-1}}{d\tau^{m-1}} g_0(\tau)$
$\sum_{n=-\infty}^{\infty} \delta(t-mn)$
×
$h_{m-1}(\tau) = \frac{\tau^{m-1} h_0(\tau)}{(m-1)!}$
$\sum$
f(t)

Fig. 2(a) System Representation for Eq. (30)



f(t)
$h_0(\tau) = (\mathrm{sinc}\,\frac{\tau}{m})^m$
$\sum_{n=-\infty}^{\infty} \delta(t-mn)$
×
$g_0(\tau) = \mathrm{sinc}\,\tau$
$h_{m-1}(\tau) = \frac{\tau^{m-1} h_0(\tau)}{(m-1)!}$
$\sum_{n=-\infty}^{\infty} \delta(t-mn)$
×
$\frac{d^{m-1}}{d\tau^{m-1}} g_0(\tau)$
$\sum$
f(t)

Fig. 2(b) System Representation for Eq. (32)

ON THE SELECTION OF OPTIMUM EXPONENTS FOR THE SERIES REPRESENTATION OF CERTAIN STOCHASTIC PROCESSES

ARTHUR J. CARLSON
University of Missouri, Columbia, Missouri

WILLIAM D. WADE
University of Iowa, Iowa City, Iowa

ABSTRACT

In the computer analysis of certain types of signals, orthonormalized exponential functions have been used by many researchers. An arbitrary finite energy signal which begins at t = 0 is easily represented as a series of specified exponential functions.

This paper deals with the selection of optimum exponents to approximate a specified statistical ensemble of signals with minimum average energy error. The ensemble is first projected onto a subspace spanned by a subset of a complete set of exponential functions. The dimensionality, D, of this initial pre-processing subspace is chosen to be large enough to represent the ensemble with negligible average energy error. Working with the ensemble as represented by the D-dimensional representation, a set of $N < D$ optimum exponents is then sought which gives an N-dimensional representation with minimum average energy error.

Formulas are developed which facilitate the numerical evaluation of the average energy error and all of its first partial derivatives with respect to the parameters of the optimum representation. A theorem is presented which simplifies the calculation of the derivatives so that a numerical search technique based upon the gradient of the error becomes practical. Results of some numerical examples are presented.

DETECTING RACES, HAZARDS AND OSCILLATIONS IN LOGICAL SYSTEMS

E. L. CALVIN
Autonetics, A Division of North American Aviation, Inc.
Anaheim, California

ABSTRACT

The fundamentals of the application of ternary algebra to logical systems are discussed in this paper. A simple interpretation of the third value introduced by ternary algebra is given and the basic Boolean operations are extended to the ternary system.

The mechanization and functional description of a computer program which uses ternary algebra and has found extensive application in the analysis of aerospace control systems are presented. The program accepts a system model in the form of Boolean equations and computes a system response using ternary algebra. A list processing technique to effect selective equation solution which results in computer time savings is presented.

In logical systems simultaneous change of two or more variables can result in malfunctions known as races and hazards. Another logical system malfunction is oscillation. Recent work has shown that these malfunctions can be detected through the use of ternary algebra. Muller[1] suggested the use of ternary algebra in the analysis of transient conditions in electronic switching circuits. Yoeli and Rinon[2] established the mathematical basis for steady state and transient analysis of combinational switching circuits using ternary algebra. Eichelberger[3] extended this to multiple input changes and to sequential circuits.

The principle features of a computer program which permit this malfunction detection are:

1. Direct operation on Boolean equations
2. Use of ternary algebra
3. Use of selective equation solution to reduce computation time
4. Detection of malfunctions caused by simultaneous changing of any number of inputs

Logical modeling is applied not only to switching circuits and discrete systems, but to continuous systems in which logical interdependencies can be defined. The digital computer makes the application feasible by reducing the time and expense required for analysis. Binary simulation programs compute a deterministic response to input state changes. The ternary simulation program computes a response which is not necessarily deterministic.

This paper provides insight into the application of ternary algebra to logical systems. A simple interpretation of the third value introduced by ternary algebra is given, and the basic Boolean operations are extended to the ternary system. The mechanization and functional description of a computer program which computes a system response using ternary algebra are presented. This program has found extensive application in the analysis of aerospace control systems.

FUNDAMENTALS

In this section, the reasons for using ternary algebra to study logical systems are investigated. It will be seen that ternary, or three-state, algebra is a natural extension of binary algebra to cover both steady state and transient conditions. Basic algorithms for evaluation of Boolean expressions are presented.

A basic precept of the application of binary algebra to switching circuits is that the signals have reached their end points. This situation precludes the possibility of investigating properly transient conditions in logical systems with binary algebra, and normally results in the non-detection of races and hazards. The application of ternary algebra to logical systems provides a technique which analytically allows for transient conditions and will automatically detect races and hazards. Eichelberger[3] has provided the mathematical basis for this.

The application of ternary algebra to logical systems involves the introduction of a third, or indeterminate state to the widely known ON and OFF states of Boolean algebra. This indeterminate or "don't know" state is intended to make the algebraic system equivalent to the actual system even during transition times. The following explanation makes this step plausible.

Ideally, the variables in a logical system can assume two and only two values. During the transient period when a variable is changing values, however, deviation from this ideal is evident. For example, in electronic circuits the "binary" transition is a continuous function rather than a discrete one. This condition is shown in Figure 1.

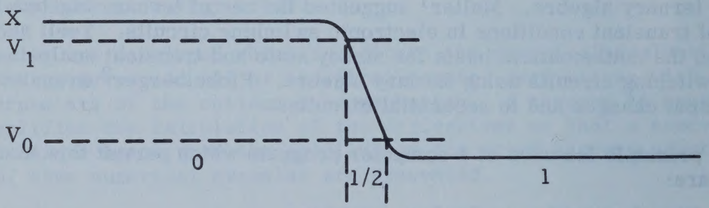


Figure 1. A Continuous Function "Binary" Transition

The variable X is considered to be a "1" when it is above V_1 and a "0" when it is below V_0. During the time the signal is between V_0 and V_1, its value is said to be indeterminate and could be either a "0" or a "1". Depending on the electrical characteristics of the gates, it could simultaneously affect one logic gate as a "1" and another as a "0".

The ternary approximation to this continuous signal (Figure 1) uses a 1/2 value to represent the signal while it is indeterminate. It should be recognized that the 1/2 state represents an indeterminable binary state, which must be either a "0" or a "1" and not necessarily a voltage which happens to lie within a given range. The choice of the value 1/2 to represent the indeterminate state permits simple arithmetic interpretations of Boolean operations on three-state variables. The machine representation of 0, 1/2 and 1 states are -1, 0, and 1, respectively. This choice still enables the use of the minimum and maximum algorithms for the AND and OR operations, while reducing the operation of complementation to simple negation.

An analogous situation exists for contact circuits and can be interpreted as follows. If a relay is initially in its non-operated state and is energized at time $t = 0$, the relay is in its operated state at time $t = T_{max}$, the maximum operate time of the relay. From this the state of the relay for $0 \leq t < T_{max}$ cannot be determined. At any particular time during this period some contacts of the relay may be operated but with a large contact resistance and some arcing. Whatever the situation, the state of the relay and its contacts is uncertain. Thus, the indeterminate state of the relay and its contacts is uncertain. Thus, the indeterminate state is attached to a relay which is in a dynamic situation. An indeterminate value has been attached to both electromechanical and electronic signals and devices during transient conditions. It should be pointed out that the indeterminate state will also be used to represent devices in steady state. The explanation is simple. Consider a circuit containing a flip-flop. When the circuit is initially energized, the state of the flip-flop is unknown. It, of course, must be in one of its two states, but until it is given a definite set or reset, the state is unknown and is referred to as being indeterminate.

With the use of a third value to represent binary signals substantiated, Boolean operations must be extended to the new algebraic system. Table 1 shows how the operations of addition (+), multiplication (·) and complementation (-) are extended to the ternary system. Note that the entries in the table are identical to those for binary determinate conditions. The table entries for which one variable is indeterminate may easily be defined by setting the indeterminate input alternately to 0 and 1. If the output is constant, the entry should be that value. If the output changes, the entry should be 1/2, or indeterminate. It can be seen that as in the binary system, the ternary AND and OR operations can be reduced to minimum and maximum functions respectively.

It is evident from the table that the Boolean identities $a \cdot \bar{a}=0$ and $a+\bar{a}=1$ are not valid in the ternary system. The arithmetic algorithm $\bar{a}=1-a$ can easily be verified and is applicable to the ternary system. This situation introduces important constraints on the implementation of ternary simulation as will be seen in the mechanization section of this paper.

Basically, the ternary simulation program operates in the following manner. When an input variable is to change states, say from 0 to 1, it is first set to the indeterminate state. The equations for the system are than solved using ternary algebra to allow the effects of the indeterminate state to be propogated throughout the logical system. The input transition is completed by setting the input variable to its final state (1) and again solving the equations. If two or more input variables are to change simultaneously, or nearly so, they are set indeterminate and then to their final states together.

A variable in the system may remain indeterminate after the completion of an input transition. This indicates that the state of the variable cannot be known with the given information, and that a race condition exists in the system. Further analysis is required to find exactly where the race condition exists for a particular input transition; the ternary simulation indicates only the subset of elements which have indeterminate states. This subset of elements is normally all of the feedback loop elements which are involved in the race, plus some of the elements dependent on them.

A logical variable may have a state which is either 0 or 1 before and after an input transition, but have a spurious 1 or 0 pulse during the transition. This defines a static hazard for that variable. In the ternary simulator this is shown by a transient indeterminate state with the before and after states being firm.

An oscillation in the logical system is another situation which will result in a set of elements remaining in the indeterminate state. All oscillating elements will be indeterminate after the input transition is complete. Further analysis is usually required to determine which elements are causing the oscillation.

Table 1. Extension of Boolean Operations to Ternary System

BINARY

TERNARY

ADDITION (+)

BINARY

		A	
		0	1
B	0	0	1
	1	1	1

TERNARY

		A		
		0	1/2	1
B	0	0	1/2	1
	1/2	1/2	1/2	1
	1	1	1	1

MULTIPLICATION (·)

BINARY

		A	
		0	1
B	0	0	0
	1	0	1

TERNARY

		A		
		0	1/2	1
B	0	0	0	0
	1/2	0	1/2	1/2
	1	0	1/2	1

COMPLEMENTATION (-)

BINARY

A	$\overline{A}$
0	1
1	0

TERNARY

A	$\overline{A}$
0	1
1/2	1/2
1	0

MECHANIZATION

Any system whose constituent components or subsystems demonstrate an interdependency resolvable to a finite number of formal logic statements is a candidate for analysis through logical modeling. The switching circuit is an example of a type of system for which the advantages of logical modeling have been extensively exploited. The development of a logical model of a system to be analyzed on a computer involves two problems. The first is the formulation of the logical relationships which represent system operation. The second problem is delineation of these relationships in a manner recognizable by the computer.

Boolean algebra is an easily understood yet effective method of specifying logical relations. In this method, the value of a dependent binary variable is equated to a logical expression of other binary variables. These binary variables, or elements, can be used to represent physical devices which have two states, e.g., a detonated or non-detonated squib, or they can be used to represent the presence or absence of abstract conditions. Boolean algebra will be used to resolve the problem of delineation of the logical relationships. Each binary element in the logical system will be given an alphanumeric name not exceeding six characters in length. These element names will be combined using the asterisk (*), plus sign (+), slash (/) and equal sign (=) to represent the Boolean operations of multiplication, addition, inversion, and equality, respectively.

The problem of the formulation of the logical relationships is considerably more difficult. Usually the system model must be developed from some basic information format, e.g., a schematic diagram, primitive connection information, or logic flow diagram. Whatever the initial format, the investigation of transient response is the aim, and it must be insured that the model accurately represents the logical system during both transient conditions and steady state conditions[1]. That is, the model must be equivalent to the actual system in the ternary sense.

As an example of the caution which must be exercised when modeling a system, consider the contact network of Figure 2.

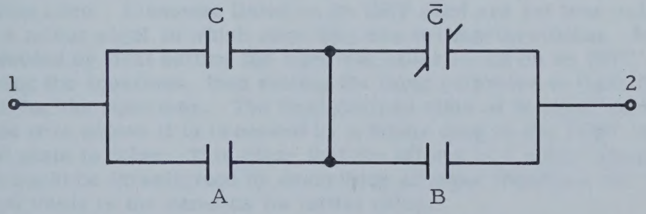


Figure 2. Sample Contact Network

When writing the transmission function between nodes 1 and 2 for binary analysis, the usual procedure is to omit the $C \cdot \bar{C}$ term from the function because $C \cdot \bar{C} = 0$ for binary variables. As was stated earlier, this identity is not valid in the ternary system. Deletion of this term from the function eliminates an existing static hazard from the model and may result in an inaccurate transient analysis of the system. Another simplification sometimes used is the reduction of $C \cdot B + A \cdot \overline{C} + A \cdot B$ to $C \cdot B + A \cdot \overline{C}$. While these expressions are equal for binary algebra, it is easy to see that if $A = 1$, $B = 1$, and $C = 1/2$, the identity is not valid for ternary algebra. Thus while Boolean notation is a good vehicle for communicating the necessary logical relationships to the computer, the identities developed within the framework of binary algebra must not be used indiscriminately.

FUNCTIONAL DESCRIPTION

The ternary simulation program described here places heavy reliance on the techniques of list programming[4]. The list structure is symmetrical, i.e., it uses both left and right linkage directors. List programming provides an easy means for storing Boolean equations and the dependency relations required for selective equation solution.

Internally, each logical element name is stored in an array A. The location in the array which contains a particular name is used to represent that name so that the machine can work with a number rather than a six-character alphanumeric name when referring to that element. Each element called out in the Boolean equations has two lists associated with it. LISTA contains the right hand portion of the Boolean equation for the element involved. LISTB is a dependency list and contains the number of each element dependent on that element. A simple example illustrates this operation. Given the set of Boolean equations

ALPHA = BETA*/A + C * TIMER

BETA = D7 + A * C * BETA

TIMER = INPUT + BETA

LISTA, LISTB, and the array A are shown in Table 2.

Table 2. LISTA, LISTB and Array A

A(1)=ALPHA	LISTA(1)	LISTA(2)	LISTA(3)	LISTA(4)	LISTA(5)	LISTA(6)	LISTA(7)
A(2)=BETA	2	6			7		
A(3)=A	-3	+			+		
A(4)=C	+	3			2		
A(5)=TIMER	4	4					
A(6)=D7	5	2					
A(7)=INPUT							

	LISTB(1)	LISTB(2)	LISTB(3)	LISTB(4)	LISTB(5)	LISTB(6)	LISTB(7)
		1	1	1	1	2	5
		2	2	2			
		5					

Note that since there are no equations for the independent elements A, C, D7, and INPUT, the corresponding equation lists for those elements are empty. LISTB(1) is empty because the dependent element ALPHA never appears on the right hand side of any equation. The element A appears in complemented form in the equation for ALPHA. The corresponding entry in LISTA(1) is -3, meaning that the value of element 3 is to be complemented.

A common method of solving a set of Boolean equations is to iterate the entire set of equations, solving every equation until the equation set is solved without anything changing. This iterating method is slow and inefficient for large equation systems. A faster and more efficient method is to use selective equation solution. If it is given that an element has changed values, only those equations affected by that element need be solved. Using the equations given in Table 2, if D7 (element number 6) changes value, LISTB(6) indicates that only the equation for element number 2 (BETA) need be solved, unless, of course, that element changes also.

If the value of BETA changes, the equations for elements 1, 2, and 5 must be solved. The equation solving portion of this program maintains a list of equations which it must solve. When the value of an element changes, the appropriate LISTB list is referenced and the elements found there are added to the bottom of the list of equations to be solved. The top entry in this list is then removed and that equation solved. When the value of the solved equation does not change, the program proceeds to the next equation to be solved. This continues until the list of equations to be solved is empty.

The equation solving portion of the program prints the name of each element that changes and its new state. It also prints out a flag message when a static hazard occurs. This is accomplished in the following manner. Two different state vectors are maintained. One vector contains the current state of each element and the other contains the most recent firm state. The current state may be true, false or indeterminate, while the firm state must be either true or false. If an element changes state, its previous value was indeterminate, and its new value is the same as its last firm state, a static hazard exists, and a message is printed.

Program control is under the direct influence of data cards which describe initial conditions and the input sequence to be executed. The four allowable input control cards are INIT, INPT, TIME, and DEND which represent initial condition specification, input transition description, delay variable operation, and end of case, respectively.

All elements are initialized to the false state unless specified otherwise by an initial condition card. Elements listed on an INIT card are set true unless they are preceded by a minus sign, in which case they are set indeterminate. Input transitions are executed by first setting the input variables listed on an INPT card indeterminate, solving the equations, then setting the input variables to their final states and again solving the equations. The final desired state of an input variable is assumed to be true unless it is preceded by a minus sign on the INPT card, in which case the final state is false. It is clear that the effects of a static hazard on a system input could be investigated by describing an input transition for that input for which the final value is the same as its initial value.

Delays often are inserted at various points in a system to insure that one event will occur before another. These delays can be simulated by representing them as an inhibit gate and listing the inhibit function on a TIME card. Variables listed on TIME cards are set indeterminate, and then true after the input transition described on the INPT cards is completed. The DEND card signals the end of a case and has options for state vector printout and initial condition reset.

CONCLUSION

The use of ternary algebra to find the response of a logical system to input changes results in a response that is non-deterministic when a critical race or oscillation is involved. Hazards are detectable by analysis of the ternary response. This makes possible the use of ternary algebra to detect malfunctions in logical systems. The ease of use provided by Boolean equation input and the speed provided by the technique of selective equation solution have helped make the ternary simulator a useful and inexpensive analysis tool.

REFERENCES

1. D. E. Muller, "Treatment of Transition Signals in Electronic Switching Circuits by Algebraic Methods," IRE Transaction on Electronic Computers, EC-8, 401 (1959).
2. M. Yoeli and S. Rinon, "Application of Ternary Algebra to the Study of Static Hazards," Journal of the ACM, Vol.11, No. 1 (January, 1964).
3. E. B. Eichelberger, "Hazard Detection in Combination and Sequential Switching Circuits," IBM Journal, (March, 1965).
4. J. Weizenbaum, "Symmetric List Processor," Communications of the ACM, Vol. 6, No. 9 (September, 1963).

GENERATION OF AN OUTPUT REGULAR EXPRESSION OF A SEQUENTIAL MACHINE WITH A SPECIFIED INPUT REGULAR EXPRESSION

S. S. YAU
Information-Processing and Control Systems Laboratory and
Department of Electrical Engineering
Northwestern University
Evanston, Illinois

ABSTRACT

It has been shown that the set of output sequences of a sequential machine due to a regular set of input sequences is regular. However, there is no method of generating the regular expression for the output regular set. The purpose of this paper is to present such a technique which is based on constructing the output graph from the state graph of the sequential machine and the input regular expression. Once the output graph is obtained, the regular expression can readily be obtained by using the signal flow graph techniques proposed by Brzozowski and McCluskey. The technique presented in this paper can be applied to both Moore and Mealy machines.

I. INTRODUCTION

A sequential machine M is a system with a finite input set $I = \{i_1, i_2, \ldots, i_u\}$, called the input alphabet, a finite output set $W = \{w_1, w_2, \ldots, w_v\}$, called the output alphabet, a finite state set $S = \{s_1, s_2, \ldots, s_r\}$, a function f, called the next-state function, such that

$$s(t+1) = f[s(t), i(t)]$$

and a function g, called the output function, such that

$$w(t) = g\ [s(t)] \qquad \text{Moore model}$$

$$w(t) = g\ [s(t), i(t)] \qquad \text{Mealy model}$$

where $s(t)$, $i(t)$, and $w(t)$ are, respectively, the present state, the present input, and the present output of M, and $s(t+1)$ denotes the next state of M. M is said to be deterministic if both f and g are single-valued. It is seen that the relation between an input symbol i_j of the input alphabet I and its corresponding output symbol w_k of the output alphabet W is determined by the next-state function f and the output function g through the present states $s(t)$ of M. For a given input sequence and the present state of M, we can find the corresponding output sequence of M. Ginsburg and Rose [1] have shown that sequential machines have the property of preserving regularity. That is, if the set of input sequences to sequential machine M is regular, then the set of corresponding output sequences produced by M must also be regular. Because English and a number of natural languages are, from a certain point of view, regular sets [2, 3], and because sequential machines and regular sets are closely related, it is natural to consider the application of sequential machines to machine translation. In order to do this, we need a technique for generating a regular expression for the set of output sequences of a given sequential machine due to a specified input regular set. However, there is, so far, no such a technique available. The purpose of this paper is to present such a tech-

---

The work reported here was jointly supported by the U.S. Air Force Office of Scientific Research under Grant No. AF-AFOSR-98-65 and the U.S. Army, Navy, and Air Force in a Joint Services Electronics Program under Contract No. N00014-66-C020A03.

nique which is based on the construction of a graph, called the output graph, from the state graph of the sequential machine and the regular expression for the input regular set. Once this graph is obtained, the regular expression for the corresponding output regular set can readily be found by using the signal flow graph techniques proposed by Brzozowski and McCluskey [4]. The method presented in this paper can be applied to both Moore and Mealy models of deterministic sequential machines.

II. REGULAR EXPRESSIONS AND SEQUENTIAL MACHINES

A set of sequences is regular if and only if it is expressible by a regular expression [5,6]. In this section, we shall briefly review the regular expressions and the relations between regular sets and sequential machines.

Before we discuss regular expressions, we would like to define some terms related to sets of sequences. Consider two sets Q_1 and Q_2 of sequences formed by the symbols of the alphabet $A=\{a_1, a_2, \ldots, a_p\}$. Q_1 and Q_2 are said to be equivalent if every sequence in Q_1 is in Q_2 and every sequence in Q_2 is also in Q_1 This relation is denoted by $Q_1=Q_2$. The union of Q_1 and Q_2, denoted by Q_1+Q_2, is the set of sequences consisting of all the sequences in either Q_1, or Q_2, or both. The concatenation of Q_1 and Q_2, denoted by $Q_1.Q_2$, or simply Q_1Q_2, is the set of sequences formed by taking any sequence in Q_1 followed by any sequence in Q_2. It is noted that $Q_1Q_2 \neq Q_2Q_1$. The concatenation of m identical sets Q_1 may be written as Q_1^m. The star of Q_1, denoted by Q_1^*, is the set of sequences represented by the expression $(\lambda+Q_1+Q_1^2+Q_1^3+\ldots)$, where λ is the null sequence or the sequence of zero length. The length of a sequence is the number of symbols from the alphabet in the sequence. Let the empty set of sequences be denoted by $\emptyset$. Then, we may define a regular expression as follows:

Given an alphabet $A=\{a_1, a_2, \ldots, a_p\}$ and the symbols λ and $\emptyset$, and the operators "+", ".", and "*", a regular expression is defined recursively as follows: 1) Each of the symbols λ, and $\emptyset$, and the symbols in A is a regular expression. 2) If R_1 and R_2 are regular expressions, so are (R_1+R_2), $(R_1.R_2)$, and R_1^*. 3) No string of symbols is a regular expression unless it is formed by 1) and 2) in a finite number of steps.

It is noted that there are usually several different regular expressions representing the same regular set of sequences. However, it is very difficult and sometimes impossible to transform from one regular expression to another even if we know they are equivalent [7]. This is, in fact, the most serious shortcoming of using regular expressions to represent regular sets of sequences.

Kleene [5] has shown that starting from any state of a sequential machine M the set of input sequences formed by the input alphabet I can cause M to produce a desirable output, which occurs at a number of states of M, if and only if the set of input sequences is regular. McNaughton and Yamada [8] have presented techniques for converting the state graph of a Moore machine to a regular expression of its input alphabet I, and vice versa. For Mealy machines, Ott and Feinstein [9] have given a method for generating the state graph from a regular expression of the input alphabet. Later, Brzozowski and McCluskey [4] have found a very efficient technique for obtaining a regular expression of the input alphabet from the state graph of a sequential machine by interpreting properly the state graph as a signal flow graph. In this paper, our technique for generating an output regular expression from an input regular expression and the state graph of a sequential machine may be divided into two stages: The first stage is to construct the output graph directly from the given input regular expression and the given state graph. The second state is to find the output regular expression from the output graph by using the method proposed by Brzozowski and McCluskey [4]. Therefore, the reader is assumed to be familiar with the results in [4].

Before we present the general procedure, let us consider the effect of the three operations "union", "concatenation", and "star", in the input regular expression on the set of corresponding output sequences. It is noted that the set of output sequences due to the union of two input sequences is obviously the union of the output sequences due to the two individual input sequences. The concatenation of two input sequences may be considered as a single input sequence, and hence its output sequence can easily be found. The set of output sequences due to the star of an input sequence is not so simple, and will be discussed separately in the next section.

III. THE SET OF OUTPUT SEQUENCES DUE TO THE STAR OF AN INPUT SEQUENCE

Before we discuss the set of output sequences due to the star of an input sequence, let us consider the following case: Let q be an input sequence of length T. The input sequence $(q)^n|n\to\infty$, denoted by $|q|$, may be considered as a periodic input sequence with the period T. Suppose the input sequence $|q|$ is applied to a sequential machine M of r states when M is in state $s_{j,0}$. If M returns to $s_{j,0}$ after receiving the first input sequence q, then M will always return to $s_{j,0}$ after receiving any number of q because M is deterministic. Consequently, the corresponding output sequence is also periodic. Suppose that M does not return to $s_{j,0}$, but reaches $s_{j,1}, s_{j,2}, \ldots, s_{j,k}$ after receiving the first, second, ..., kth input sequence q respectively, where $s_{j,0}, s_{j,1}, \ldots, s_{j,k}$ are all distinct. It is clear that k cannot be greater than r-1. Hence, after receiving a maximum of r input sequences q, M has to reach one of the states $s_{j,0}, s_{j,1}, \ldots, s_{j,k}$. Let this state be $s_{j,\alpha}$, where $0 \le \alpha \le k$. Then, M will always reach the states $s_{j,\alpha}, s_{j,\alpha+1}, \ldots, s_{j,k}$ after receiving $\alpha+m(k-\alpha+1)$th, $\alpha+m(k-\alpha+2)$th, ..., $\alpha+m(2k-2\alpha+1)$th input sequence respectively, where m = 1,2, ... This indicates that the state transitions due to a periodic input sequence with a period T always form a cycle after receiving ρ input symbols, where $\rho \le (r-1)T$. Consequently, the output sequence due to a periodic input sequence is also periodic after at most ρ symbols.

Based on the above discussion, we can construct a graph G representing the output sequence of M due to a periodic input sequence $|q|$ from the state graph of M as follows: Each of the states $s_{j,0}, s_{j,1}, \ldots, s_{j,k}$ corresponding to a vertex in G. There is a branch directed from $s_{j,m}$ to $s_{j,m+1}$, m=0, 1, ... , k-1, and a branch directed from $s_{j,k}$ to $s_{j,\alpha}$. The weight on a branch in G is the output sequence of M when M moves from the beginning state to the terminal state of the branch due to the input sequence q. It follows from the above discussion that the graph G for a periodic input sequence $|q|$ contains one and only one cycle. To demonstrate this procedure, let us consider the sequential machine M_1, whose state graph is shown in Fig. 1. Suppose the periodic input sequence $|i_4 i_2 i_3|$ is applied to M_1 when M_1 is in state s_1. According to our procedure, the graph G shown in Fig. 2 is obtained. It is readily found from G that the corresponding output sequence is $w_4w_3w_1|w_1w_4w_2w_2w_3w_1|$.

Now, we are ready to draw the output graph G_W with respect to $(q)^*$. First, we define that the output of M due to the input λ is always λ. Then, it follows from the definition of $(q)^*$ that the output graph G_W with respect to $(q)^*$ is the same as the graph G for the input sequence $|q|$, except that a directed branch with the weight λ is added from the vertex $s_{j,m}$ to a new vertex $s_{j,m}^{(1)}$ for m=0, 1, ... , k, and that a directed branch with the weight λ is added from a new vertex $s_{j,0}^{(0)}$ to a vertex $s_{j,0}$ if the vertex $s_{j,0}$ has incoming branches. Thus, in the G_W with respect to

$(q)^*$, there is one vertex $s_{j,0}$ or $s_{j,0}^{(0)}$ with outgoing branches only and k+1 vertices with incoming branches only. Hence, by using the techniques proposed by Brzozowski and McCluskey [4] a regular expression representing the set of all output sequences with respect to $(q)^*$ can be found by considering G_W as a signal flow graph with $s_{j,0}$ or $s_{j,0}^{(0)}$ as the source and $s_{j,0}^{(1)}, s_{j,1}^{(1)}, \ldots, s_{j,k}^{(1)}$ as the sinks. For the above example, the output graph G_W of M with respect to $(i_4i_2i_3)^*$ for M_1 at s_1 is shown in Fig. 3, and the set of all the output sequences of M_1 due to $(i_4i_2i_3)^*$ is represented by the regular expression

$$\lambda + w_4w_3w_1(w_1w_4w_2w_2w_3w_1)^* + w_4w_3w_1(w_1w_4w_2w_2w_3w_1)^*w_1w_4w_2,$$

which can be written in the following form

$$\lambda + w_4w_3w_1(w_1w_4w_2w_2w_3w_1)^*(\lambda + w_1w_4w_2).$$

IV. THE GENERAL PROCEDURE

It follows from the discussion in the last section that the concatenation of a number of input symbols without involving any other operations in an input regular expression R_I may be considered as a unit, called the *unit input subsequence*, and the weight on each branch of the output graph G_W is the output subsequence due to a unit input subsequence only. We shall call a unit input subsequence an *initial unit input subsequence* if it appears as the unit input subsequence at the beginning of one of the input sequences in R_I. Similarly, we shall call a unit input subsequence a *terminal unit input subsequence* if it appears as the unit input subsequence at the end of one of the input subsequences in R_I. For example, we decompose the input regular expression

$$R_i = i_1i_2i_4(i_3i_1 + i_4i_2{}^*i_3)^*i_2$$

into 6 unit input subsequences, and we may write R_I in the form

$$(i_1i_2i_4)_1[(i_3i_1)_2 + (i_4)_3\ (i_2)_4{}^*(i_3)_5]^*\ (i_2)_6.$$

It is obvious that $(i_1i_2i_4)_1$ is the only initial unit input subsequence and $(i_2)_6$ is the only terminal unit input subsequence. From R_I we can also find all the order pairs of unit input subsequences each of whose concatenations is permissible in R_I. These ordered pairs are called the *permissible ordered pairs*. For the above R_I, we have the following permissible ordered pairs (unit input subsequences are represented by their subscripts): (1,2), (1,3), (1,6), (2,2), (2,3), (2,6), (3,4), (3,5), (4,4), (4,5), (5,2), (5,3), (5,6). We shall say "a unit input subsequence q_j has the set σ of permissible ordered pairs" if every permissible ordered pair in the set σ has q_j as its first element. Every unit input subsequence, except the terminal ones, should have a nonempty set of permissible ordered pairs.

After decomposing the input regular expression R_I and obtaining all the permissible ordered pairs, we can construct the output graph G_W by the following method: Draw a vertex corresponding to the starting state $s_{0,0}$ of M, and from this vertex draw a directed branch $b_{1,\alpha}$ for each initial unit input subsequence $q_{1,\alpha}$ with the corresponding output subsequence as its weight. The terminal vertex of $b_{1,\alpha}$ corresponds to the final state $s_{1,\alpha}$ reached by M after receiving $q_{1,\alpha}$. Suppose that we have two initial unit input subsequences $q_{1,1}$ and $q_{1,2}$ with the final states $s_{1,1}$ and $s_{1,2}$. Then, the vertices corresponding to $s_{1,1}$ and $s_{1,2}$ can coalesce together if and only if they satisfy the following two conditions: (a) $s_{1,1}$ and $s_{1,2}$ are the same state of M, and (b) $q_{1,1}$ and $q_{1,2}$ have the identical set of permissible ordered pairs. If the vertices for $s_{1,1}$ and $s_{1,2}$ coalesce, the parallel branches $b_{1,1}$ and $b_{1,2}$ are combined to a single branch with the union of the weights of $b_{1,1}$ and $b_{1,2}$ as its weight. Next, we repeat the same process to each of the final states of the initial unit input subsequences. For example, we repeat the process to the final state $s_{1,\alpha}$ of the initial unit input subsequence $q_{1,\alpha}$ by considering the unit input subsequences, each of which is the second element in a permissible ordered pair of $q_{1,\alpha}$, as the initial unit input subsequences. In order to reduce the complexity of the output graph G_W, the two vertices for two final states always coalesce together whenever they satisfy the conditions (a) and (b), no matter whether the two vertices are both newly derived or one newly derived and one derived before. It is noted that this process must terminate after a finite number of steps because R_I has a finite number of unit input subsequences and M is a finite-state sequential machine. After completing this process, draw a directed branch with the weight λ from each of the vertices, which is not a sink but represents a final state of a terminal unit input subsequence, to a new vertex so that the new vertex becomes a sink. Similarly, if the vertex representing the starting state $s_{0,0}$ is not a source, draw a directed branch with the weight λ from a new vertex to this vertex so that the new vertex becomes the source. The resultant graph is the desired output graph G_W of M with respect to R_I.

Theorem: A sequence is an output sequence of a sequential machine M due to an input sequence in an input regular expression R_I if and only if it can be obtained by concatenating the weights of all the branches in a branch sequence from the source to one of the sinks in the output graph G_W constructed from R_I and the state graph of M by the above method.

Proof: To prove the "only if" part, we assume that there exists an output sequence of M due to an input sequence q_j in R_I which cannot be obtained by concatenating the weights of all the branches in any branch sequence from the source to any of the sinks in the G_W. Because q_j is in R_I, q_j must be a concatenation of the unit input subsequences which is permissible in R_I. Due to the way we construct G_W, every possible concatenation of the output subsequences due to the unit input subsequences which is permissible in the R_I appears as the concatenation of the weights of the branches in a branch sequence from the source to a sink in G_W. Hence, the output sequence due to q_j must be a concatenation of the weights of the branches in a branch sequence from the source to a sink in G_W. This contradicts the assumption, and hence completes the proof of the "only if" part of the theorem.

To prove the "if" part of the theorem, we assume that there exists a concatenation of the weights of all the branches in a branch sequence from the source to a sink in G_W which is not an output sequence of M due to an input sequence in R_I. Let this branch sequence be $b_{1,\alpha}b_{2,\alpha}\cdots b_{e,\alpha}$. Then, because of the way we construct G_W, the input subsequences for $b_{1,\alpha}$ (or $b_{2,\alpha}$ when the weight of $b_{1,\alpha}$ is λ) and $b_{e,\alpha}$(or $b_{e-1,\alpha}$ when the weight of $b_{e,\alpha}$ is λ) must be the initial and terminal unit input subsequences in R_I respectively, and every pair of input subsequences for any two successive branches in this branch sequence must be a permissible ordered pair in R_I. Hence, the input sequence formed by the concatenation of the input subsequences for the branches $b_{1,\alpha}$, $b_{2,\alpha}$, ... , $b_{e,\alpha}$ must be an input sequence in R_I, and consequently the concatenation of the weights of $b_{1,\alpha}$, $b_{2,\alpha}$, ... , $b_{e,\alpha}$ must be an output sequence of M due to an input sequence in R_I. A contradiction results, and hence the theorem.

Now, our general procedure for generating an output regular expression R_W of a sequential machine M with a specified input regular expression R_I may be summarized as follows:

1) Decompose R_I into unit input subsequences, and find all the initial and terminal unit input subsequences and permissible ordered pairs.

2)Construct the output graph G_W from the state graph of M and the information obtained in 1) as described in the construction method.

3)By using the signal flow graph techniques proposed by Brzozowski and McCluskey [4] , find a regular expression representing the sequences which are the concatenation of the weights of the branches in all possible branch sequences from the source to each sink in G_W. Then, the desired R_W is the union of the regular expressions for all the sinks.

To demonstrate this procedure, let us consider the sequential machine M_2 whose state graph is shown in Fig. 4, and the input regular expression

$$R_I = (i_1i_2 + i_2i_1)^*i_2.$$

Suppose the starting state of M_2 is s_1. We like to find a corresponding output regular expression R_W. First, we decompose R_I into the unit input subsequences as follows:

$$R_I = [(i_1i_2)_1 + (i_2i_1)_2]^*(i_2)_3$$

All the three unit input subsequences are initial unit input subsequences and only $(i_2)_3$ is a terminal unit input subsequence. The permissible ordered pairs (unit input subsequences are represented by their subscripts) are the following: (1,1), (1,2), (1,3), (2,1), (2,2), (2,3). Then, according to our construction method for G_W, we obtain the G_W shown in Fig. 5. It is noted that the vertex in G_W for the starting state is not a source, and a branch with the weight λ has been added so that the new vertex becomes the source for G_W. By considering the G_W as a signal flow graph [4] , we obtain the following output regular expression R_W by inspection:

$$[(w_1w_4+w_3w_1)(w_2w_1)^*w_4w_2(w_1w_2)^*w_3w_3]^*w_3$$
$$+(w_1w_4+w_3w_1)[w_2w_1+w_4w_2(w_1w_2)^*w_3w_3(w_1w_4+w_3w_1)]^*w_4$$
$$+ (w_1w_4+w_3w_1)(w_2w_1)^*w_4w_2[w_1w_2+w_3w_3(w_1w_4+w_3w_1)(w_2w_1)^*w_4w_2]^*w_1.$$

It is seen that in the above example the most complicated part in constructing G_W is due to the star over the union operation in R_I. In fact, the construction of G_W becomes even more complicated if R_I involves the star over the star and union operations. For instance, the output graph G_W of M_2 at s_1 with respect to the input regular expression $(i_1i_2^* + i_2i_1)^*$ is shown in Fig. 6, which is more complicated than the G_W shown in Fig. 5. This situation is not surprising because $(i_1i_2^* + i_2i_1)^*$ consists of more varieties of input sequences than $(i_1i_2 + i_2i_1)^*i_2$.

V. CONCLUSIONS AND FURTHER PROBLEMS

In this paper we have presented a technique for generating an output regular expression of a sequential machine with a specified input regular expression. This technique can be applied to both Moore and Mealy machines. It is seen from the examples that the output regular expressions obtained by this method are usually complicated and can hardly be simplified even for quite simple output graphs. This is, in fact, one of the characteristics of using regular expressions to represent regular sets of sequences.

To do the inverse, this technique can easily be extended to find an input regular expression for a sequential machine M so that M produces a specified regular set of output sequences, provided that the specified regular set of output sequences is a subset of the set of all possible output sequences produced by M. Because it has been shown [10] that it is recursively solvable to determine whether or not there exists a sequential machine M so that M produces a specified output regular expression R_W after receiving a specified input regular expression R_I, a further problem along this line of research is to find a method for generating a state graph for M for a specified realizable pair (R_I, R_W).

ACKNOWLEDGMENT

The author wishes to thank Dr. K. C Wang for helpful discussions at the early stage of this work.

REFERENCES

[1] S. Ginsburg and G. F. Rose, "Operations which preserve definability in languages," Jour. Assoc. Computing Machinery, vol. 10, pp. 175-195; 1963.
[2] V. Yngve, "A model and a hypothesis for language structure," Proc. Am Philos. Soc., vol. 104; pp. 444-466; 1960.
[3] V. Yngve, "Computer programs for translation," Scientific American, vol. 204, pp. 68-76; 1962.
[4] J. A. Brzozowski and E. J. McCluskey, "Signal flow graph techniques for sequential circuit state diagrams," IEEE Trans. on Electronic Computers, vol. EC-12, pp. 67-76; 1963.
[5] S. C. Kleene, "Representation of events in nerve nets and finite automata," Automata Studies, Annals of Math. Studies, No. 34, Princeton Univ. Press, Princeton, N. J., pp. 3-41; 1956.
[6] I. M. Copi, C. C. Elgot and J. B. Wright, "Realization of events of logical nets," Jour. Assoc. Computing Machinery, vol. 5, pp. 181-196; 1958.
[7] H. Ghiron, "Rules to manipulate regular expressions of finite automata," IRE Trans. on Electronic Computers, vol. EC-11, pp. 574-575; 1962.
[8] R. McNaughton and H. Yamada, "Regular expressions and state graphs for automata," IRE Trans. on Electronic Computers, vol. EC-9, pp. 39-47; 1960.
[9] G. Ott and N. H. Feinstein, "Design of sequential machines from their regular expressions," Jour. Assoc. Computing Machinery, vol. 8, pp. 585-600; 1961.
[10] S. Ginsburg and T. N. Hibbard, "Solvability of machine mappings of regular sets to regular sets," Jour. Assoc. Computing Machinery, vol. 11, pp. 302-312; 1964.

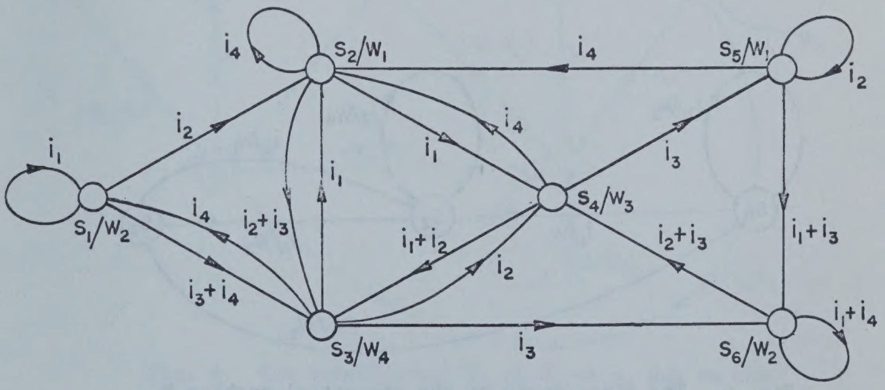


Fig. 1. The state graph of the sequential machine M_1.

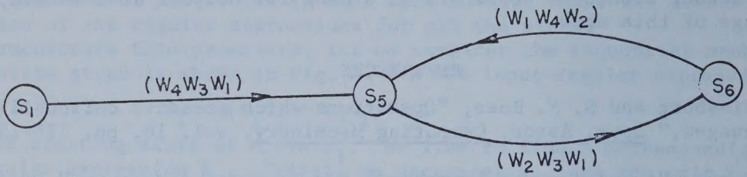

(W1 W4 W2)
(W4 W3 W1)
S1
S5
S6
(W2 W3 W1)

Fig. 2. The output graph G of M_1 at s_1 due to the periodic input sequence $\lceil i_4 i_2 i_3 \rceil$.

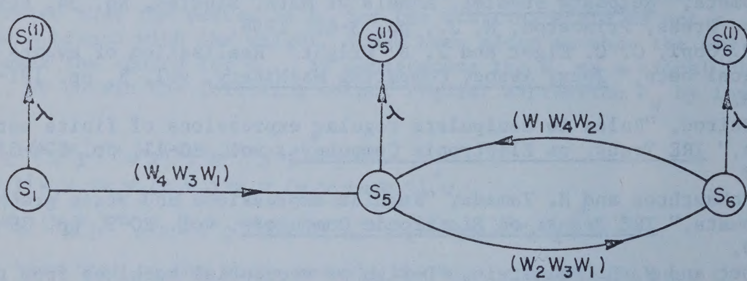

S1(1)
S5(1)
S6(1)
λ
λ
λ
(W1 W4 W2)
(W4 W3 W1)
S1
S5
S6
(W2 W3 W1)

Fig. 3. The output graph G_W of M_1 at s_1 due to the input regular expression $(i_4 i_2 i_3)^*$.

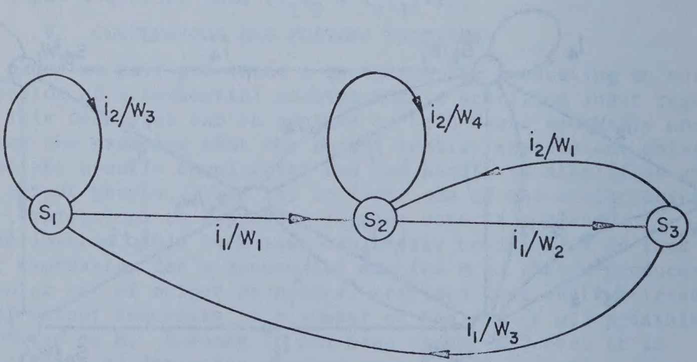

i2/W3
i2/W4
i2/W1
S1
S2
S3
i1/W1
i1/W2
i1/W3

Fig. 4. The state graph of the sequential machine M_2.

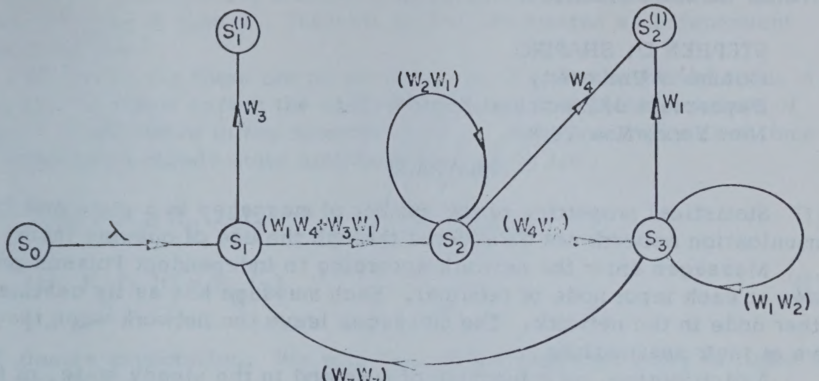

S₁⁽¹⁾
W₃
(W₂W₁)
W₄
S₂⁽¹⁾
W₁
S₀
λ
S₁
(W₁W₄+W₃W₁)
S₂
(W₄W₁)
S₃
(W₁W₂)
(W₃W₃)

Fig. 5. The output graph G_W of M_2 at s_1 due to the input regular expression $(i_1i_2+i_2i_1)^*i_2$.

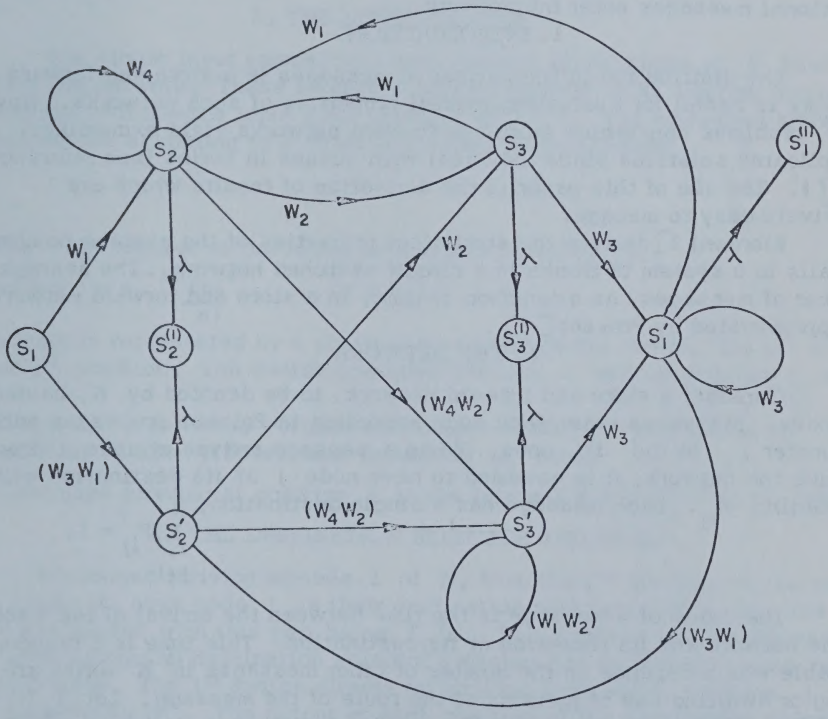

W₁
W₄
W₁
S₂
S₃
S₁⁽¹⁾
W₂
W₁
λ
W₂
λ
W₃
λ
S₁
S₂⁽¹⁾
S₃⁽¹⁾
S₁′
W₃
(W₄W₂)
λ
λ
W₃
(W₃W₁)
S₂′
(W₄W₂)
S₃′
(W₁W₂)
(W₃W₁)

Fig. 6. The output graph G_W of M_2 at s_1 due to the input regular expression $(i_1i_2{}^*+i_2i_1)^*$.

A QUEUING ANALYSIS OF THE NUMBER OF MESSAGES IN A STORE AND FORWARD COMMUNICATION NETWORK

STEPHEN D. SHAPIRO
Columbia University
Department of Electrical Engineering
New York, New York

ABSTRACT

Statistical properties of the number of messages in a store and forward communication network are considered through the use of queuing theory.

Messages enter the network according to independent Poisson processes at each input node or terminal. Each message has as its destination another node in the network. The messages leave the network when they arrive at their destinations.

A distribution, as a function of time and in the steady state, of the number of messages in the network is presented assuming there are initially no messages in the network. An examination is made of the time required to clear the network of all messages in transmission, and a lower bound is obtained on the steady state probability that all messages in the network reach their destinations within a specified time interval, during which no additional messages enter the network.

1. INTRODUCTION

The distribution of the number of messages in a store and forward network is useful for evaluating overall properties of such networks. However, problems concerning store and forward networks yield exceedingly complicated solutions since they deal with queues in series (see reference [1]). The aim of this paper is the derivation of results which are relatively easy to manage.

Riordan[2] derives the statistical properties of the average number of calls in a system of trunks in a circuit switched network. The average number of messages, as a function of time, in a store and forward network is approximated by Prosser[3].

2. THE NETWORK

Consider a store and forward network, to be denoted by N, having n nodes. Messages enter each node according to Poisson processes with parameter λ_i at the i^{th} node. When a message arrives at node i from outside the network, it is assumed to have node j as its destination with probability P_{ij}. Each message has a single destination,

$$\sum_{j=1}^{n} P_{ij} = 1.$$

The delay of a message is the time between the arrival of the message at the network and its reception at its destination. This time is a random variable which depends on the number of other messages in N which are using or awaiting use of portions of the route of the message. Let $T_{ij}(t)$ = the delay of a message entering N at node i at time t and having node j as its destination.

The key assumption is that the distribution of the delays may be adequately described by giving the distribution functions

$P\{T_{ij}(t) \leq x\} = H_{ij}(x|t)$ for all i and j. These distribution functions take into account the cumulative influences of the other messages in N on a message arriving at time t. Thus the delays are treated as independent random variables.

At $t = 0$, say there are no messages in N. As the arrival time, t, increases, we would expect the delay to increase due to the presence of messages which arrive in the interval up to t. As $t \to \infty$, we will assume each delay has a steady state distribution $H_{ij}(x)$, i.e.

$$H_{ij}(x|t_1) \geq H_{ij}(x|t_2) \qquad \text{if} \quad t_2 \geq t_1 \tag{1}$$

$$\lim_{t \to \infty} H_{ij}(x|t) = H_{ij}(x) \tag{2}$$

Let E denote expectation. We will assume the expected delay is finite for all t, $E(T_{ij}(t)) < \infty$.

Denote by $M(t)$ the number of messages in N at time t. We shall find the distribution of $M(t)$ by considering the delay process as a service time, which depends on the message arrival time, of an infinite server queue with a single input (see reference [4]). Hence $M(t)$ becomes the queue size at time t.

3. THE INPUT PROCESS

The single input corresponds to messages which arrive at N from outside the network. These messages may be thought of as arriving at an n-position switch—the input node selector of Fig. 1. The messages arrive at this switch according to a Poisson process with parameter

$$\lambda = \sum_{i=1}^{n} \lambda_i .$$

If $A(t)$ = the number of messages arriving at N in $[o,t]$,

$$P\{A(t) = a\} = e^{-\lambda t} \frac{(\lambda t)^a}{a!} \qquad a = 0,1,2,\ldots \quad .$$

Each node is represented by a position (terminal) of the switch, the i^{th} node by the i^{th} position. The switch occupies position i with probability $\frac{\lambda_i}{\lambda}$ at each arrival. It is easily verified that output at the i^{th} position is an independent Poisson process with parameter λ_i since

$P\{$a message arriving at position i in the interval $[t, t+\Delta t]\} = (\lambda \Delta t + o(t))\left(\frac{\lambda_i}{\lambda}\right)$.

4. THE DESTINATION SELECTOR AND DELAY

Messages arriving at node i of N, from the i^{th} position of the input node selector, have node j as their destination with probability P_{ij}, the delay associated with this pair being $T_{ij}(t)$. This process may be represented by a bank of random destination selector switches shown in Fig. 1; one switch for each node. The i^{th} switch directs an arriving message to position j with probability P_{ij}. The output at each position of the destination selector switch is directed to a delay box where it is held for a time $T_{ij}(t)$ before leaving the network.

Distribution of the delay

Let $T(t)$ = the delay of an arbitrary message arriving at N at t. From the

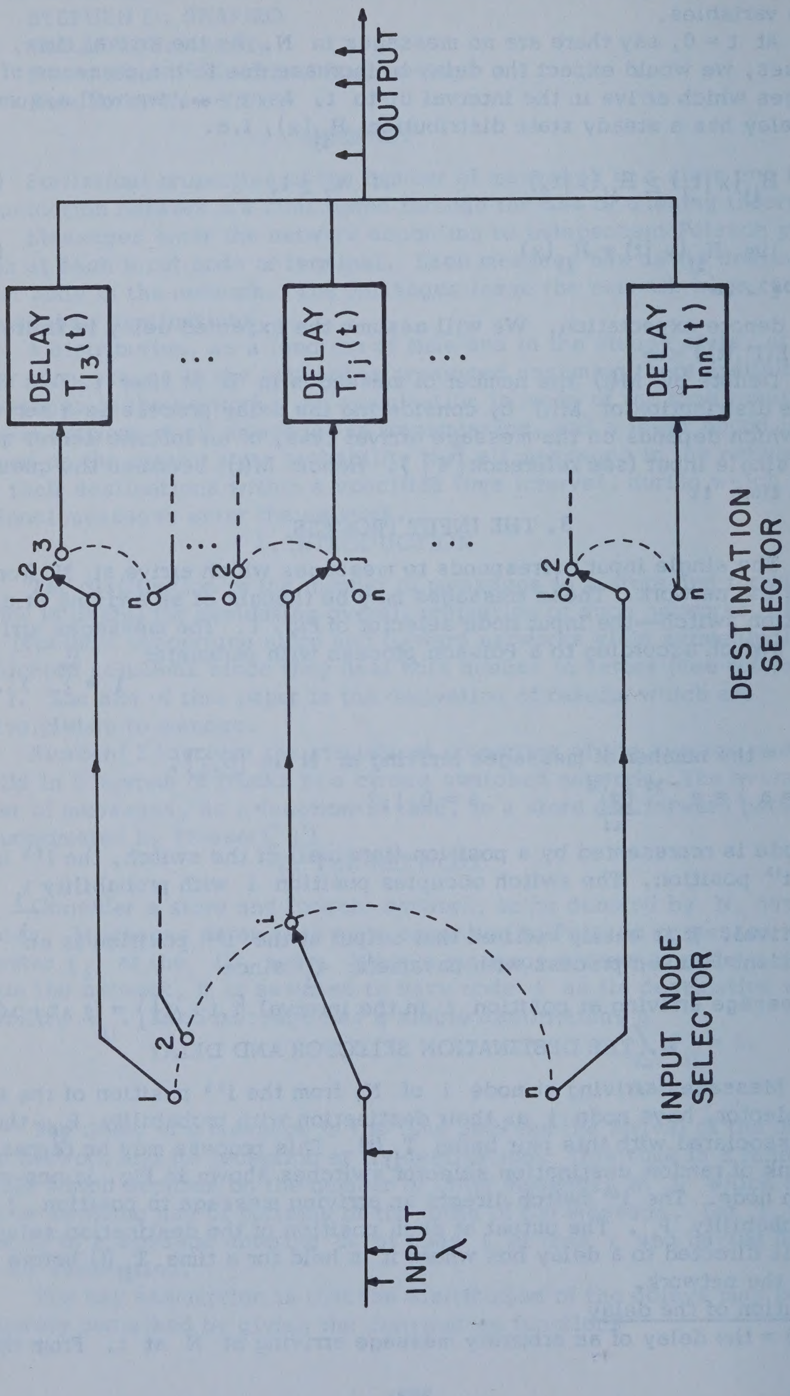

INPUT
λ
INPUT NODE
SELECTOR
1
2
i
n
DESTINATION
SELECTOR
1
2
3
n
j
DELAY
$T_{13}(t)$
DELAY
$T_{ij}(t)$
DELAY
$T_{nn}(t)$
OUTPUT

FIG. 1

discussion above and Fig. 1, we see

$$P\{T(t) \le x\} = H(x|t) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\lambda_i}{\lambda} P_{ij} H_{ij}(x|t).$$

In the steady state, from (2),

$$\lim_{t\to\infty} H(x|t) = H(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\lambda_i}{\lambda} P_{ij} H_{ij}(x).$$

In addition,

$$E(T(t)) = \int_0^{\infty} (1-H(x|t))\,dx \le \infty \text{ for all } t. \qquad (3)$$

5. THE NUMBER OF MESSAGES IN N

$M(t)$ corresponds to the queue size at time t of an infinite server queue. $T(t)$ is the service time for an arrival at time t. Arrivals are Poisson with parameter λ. We shall outline the derivation of $P\{M(t) = k\}$; extending the result in Takács, [4], to the case in which the service time depends on the arrival time, τ, and $A(t)$.

$$\text{Let } H(x|\tau, A(t) = a\} = P\{T(\tau) \le x \mid A(t) = a, 0 \le \tau \le t\}. \qquad (4)$$

Then,

$$P\{M(t) = k\} = \sum_{a=k}^{\infty} P\{M(t) = k \mid A(t) = a\}\ P\{A(t) = a\}$$

$$P\{M(t) = k \mid A(t) = a\} = \binom{a}{k} P(t,a)^k (1 - P(t,a))^{a-k} \qquad (5)$$

where $P(t,a) = P\{$ a message arriving at N in $[0,t]$ is still in N at time $t \mid A(t) = a\}$. Since we are given $\underline{a}$ Poisson arrivals, the distribution of the $\underline{a}$ arrivals is uniform in $[0,t]$; therefore,

P{ a message arrives in the interval $[u, u + \Delta u]\} = \frac{\Delta u}{t}$ for $0 \le u \le t$. From (4),

P{ a message is still N at $t \mid$ the message arrived at time u and $A(t) = a\} = 1 - H(t-u|u, A(t) = a)$.

Hence,

$$P(t,a) = \frac{1}{t}\int_0^t (1 - H(t-u \mid u, A(t) = a))\,du.$$

Since $P(t,a)$ depends on $\underline{a}$, the sum (5) cannot be closed unless more is known about the form of $H(x|\tau, A(t) = a)$. In order to close the sum and obtain a reasonable solution, let us replace $H(x|\tau, A(t) = a)$ by its average value:

$$H(x|\tau,t) = \sum_{a=0}^{\infty} H(x|\tau, A(t) = a) e^{-\lambda t} \frac{(\lambda t)^a}{a!}$$

Note, however, that t in $H(x|\tau,t)$ bounds τ from above and its value does not influence the distribution of $T(\tau)$ for $\tau \leq t$. Hence we let $t = \infty$ yielding $H(x|\tau)$. The expression for the average $P(t,a)$, in which $H(x|\tau)$ replaces $H(x|\tau, A(t) = a)$, is then given by

$$P(t) = \frac{1}{t}\int_0^t (1 - H(t-u|u))\, du$$

By using this average value, we close the sum (5):

$$P\{M(t) = k\} \cong e^{-\lambda t P(t)} \frac{(\lambda t P(t))^k}{k!} \qquad k = 0,1,2,\ldots$$

Steady State Distribution of $M(t)$

In the steady state let M = the number of messages in N; for the average $P(t)$:

$$P\{M = k\} \cong e^{-\lambda E(T)} \frac{(\lambda E(T))^k}{k!} \qquad k = 0,1,2,\ldots$$

where $T = T(\infty)$.

The relation follows since $\lim_{t\to\infty} t\,P(t) = E(T)$. This may be shown directly by making a change of variables $v = t - u$ in the expression for $P(t)$ and noting

$$\lim_{t\to\infty} tP(t) = \lim_{w\to\infty}\lim_{t\to\infty}\int_0^w (1 - H(v|t-v))dv + \lim_{w\to\infty}\lim_{t\to\infty}\int_w^t (1 - H(v|t-v))\, dv$$

The left hand integral approaches $E(T)$, by (2) and (3), providing the integration and limit with respect to t may be interchanged. The integrand of the right hand integral is bounded by $1 - H(x)$ because of (1); hence, the integral goes to zero since $E(T) < \infty$.

We note that $E(T) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\lambda_i}{\lambda} P_{ij} E(T_{ij})$ and that

$$E(M) = \lambda E(T) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i P_{ij} E(T_{ij})$$

= the expected number of arrivals during an interval equal to the expected delay.

6. TIME REQUIRED TO CLEAR N OF MESSAGES

Suppose at time t we close all the input terminals of the network. Those messages already in N continue toward their destinations. We may now ask: How much time must elapse after t until all of these messages reach their destinations? This is the time required to clear N of all messages in transmission.

Since the network is store and forward, the delay of the messages in N at t will be reduced when additional messages are prevented from entering N. Consequently, a lower bound on the probability that N is cleared after an interval of length I may be obtained by assuming that messages continue to enter N after t and that the messages in transmission do not undergo the decrease in delay. We then have in steady state, using the average value of $P(t,a)$,

$$\lim_{t \to \infty} P\{\text{ N is cleared in an interval length I }\}$$

$$\leq \sum_{k=0}^{\infty} H^*(I)^k \, P\{M = k\}$$

where

$$H^*(I) = \frac{1}{E(T)} \int_0^I (1 - H(v))\, dv .$$

This probability may be found, (see [4]), by introducing the random variables $x_1(t), x_2(t), \ldots, x_k(t)$ where $x_i(t)$ = the time, as measured from t, which is needed for the i^{th} message to reach its destination. If there are k messages in N at t,

$$P\{\text{ N is cleared after an interval of length } I \mid M(t) = k\} =$$

$$P\{x_1(t) \leq I,\ x_2(t) \leq I, \ldots,\ x_k(t) \leq I \mid M(t) = k\}.$$

Taking the limit as $t \to \infty$, this probability is greater than or equal to $H^*(I)^k$.

7. CONCLUSION

Overall properties of the store and forward network have been considered through the use of a queuing model. It is possible to extend these results to include multiple address messages by using the state assignment and general definition of delay (transmission time) given in [5].

ACKNOWLEDGEMENT

The author wishes to thank Professor R. J. Schwarz of Columbia University for his suggestions and comments.

REFERENCES

[1] L. Kleinrock, Communication Nets, New York: McGraw-Hill, 1964.

[2] J. Riordan, "Telephone Traffic Time Averages," Bell System Technical Journal, vol. 30, no. 4, October, 1951.

[3] R. T. Prosser, "Routing Procedures in Communication Networks, Part I: Random Procedures," IRE Transactions on Communication Systems, vol. CS-10, pp. 322-329, December, 1962.

[4] L. Takács, Introduction to the Theory of Queus, New York: Oxford University Press, 1962.

[5] S. D. Shapiro, "Random Store and Forward Communication Networks", Proceedings of the Symposium on Generalized Networks, New York: Polytechnic Institute of Brooklyn, 1966. (to appear)

TOPOLOGICAL STRUCTURES OF INFORMATION RETRIEVAL SYSTEMS*

R. T. Chien and F. P. Preparata
Coordinated Science Laboratory and Department of Electrical Engineering
University of Illinois
Urbana, Illinois

ABSTRACT

This paper considers the problem of information retrieval from the point of view of graph theory. In this formulation documents are represented as nodes and relationships among the documents are represented by edges. Two types of graphs are introduced, namely the similarity graph which is based on subject-content correlation and the citation graph, which is derived from direct citation linkages among documents. Several distance measures are considered and evaluated with regard to retrieval operations.

I. INTRODUCTION

Within the scope of this paper we shall consider an information retrieval system to consist of two major components, namely, a document collection and a retrieval procedure, that is, a systematic way of selecting a subset of documents of the collection according to a given criterion.

The documents in the collection are coupled to one another in many different respects, such as subject content, form, authorship, citations, etc. Two of these facets, namely subject content and citations, have been exploited for application in retrieval.

In a great many modern information retrieval systems the characteristics in subject content are expressed in terms of subject descriptors. Attached to each document is a set of subject descriptors which characterizes the subject content of the document. A measure of the similarity between a pair of documents can then be obtained by comparing their assigned descriptors. Characterizations of documents through the use of subject descriptors is known as coordinate indexing.

In retrieval operation a query is presented to the system which describes a profile of the type of documents to be retrieved from the collection. In most systems employing coordinate indexing today the query is given in terms of a set of descriptors or some logical function thereof. For instance, we may ask for all documents that deal with the "decoding" of "Bose-Chandhuri-Hocquenghem Codes" that are published in the "Transactions of IEEE on Information Theory" since "1964", where those terms under quotation signs are descriptors.

Another type of retrieval systems are based on citation indexing. In this type of systems citation information among documents is stored in the system. The query is given in terms of specifying accession documents in the network. For instance, one might wish to retrieve all documents citing a document d or one might wish to retrieve all documents that are cited by document d. Retrieval operations based on multi-generation citations are theoretically feasible but so far have not received much attention.

---

*This work is supported in part by the National Science Foundation Grant NSF GF-690 and in part by the Joint Services Electronics Programs (U. S. Army, U. S. Navy and U. S. Air Force) under Contract No. DA 28 043 AMC 00073(E)

In comparing the two popular schemes, citation indexing is easy to instrument but is limited in scope in that it derives information only from existing direct linkages in the document collection. This restriction is reflected in the usual incompleteness of retrieval results when one is interested in searches based on subject content.

On the other hand, coordinate indexing works well only if the indexed document collection is relatively homogeneous and the query well-defined. For requests from research scientists the query is always aimed at the intersection or the union of several narrow and ill-defined disciplines. As a result, the outcome is usually contaminated with large amounts of irrelevant material.

Aimed at retrieval procedures that will produce sharper and more complete responses we propose the study of potential systems that combine the resources of both the coordinate-indexing approach and the citation methods. To minimize the inconsistency between indexing and retrieval we choose to represent all queries in terms of documents. To state it formally, the problem treated in this paper is one of finding an information retrieval system that combines the advantages of both the coordinate indexing and citation indexing. A typical retrieval operation would be the retrieval of a set of documents that is "close" in some reasonable measure to a given document profile. To facilitate instrumentation emphasis is placed on easily-implemented systems.

II. THE CORRELATION GRAPH

The main consideration in this section will be document couplings that are subject-content based. Although a number of studies have been made in this area involving fairly complicated couplings and their interactions, the type of couplings to be investigated here will be relatively simple in nature as our chief objective dwells on the question of optimum combination of subject-content based indexing and non-subject-content based indexing.

Let us consider a coordinate indexing scheme in which each document is assigned a number of descriptors. For a typical system the total number of descriptors will be of the order of 10,000 while each document may be assigned ten to fifteen descriptors on the average. A typical curve for descriptor frequency is given in Figure 1. The behavior of the curve sketched in Figure 1 can be explained as follows. It is observed that typically there are two kinds of descriptors. Descriptors of the first kind may be termed general descriptors and have a high probability of being used for many documents. Descriptors of the second kind are specialized in nature and have a low probability of being used but provide the system with a tremendous amount of selectivity whenever they are present.

The dichotomy of the descriptor population points up the difficulty in indexing resolution. In the interest of efficiency it is necessary to keep the number of descriptors, especially descriptors of the general type, small. The thesaurus of any practical system is therefore usually the result of compromises. While the initial resolution may be adequate for the initial collection and subject to most queries, the system may not perform satisfactorily when the document collection grows or when the system can not be defined clearly with the system's limited vocabulary.

As pointed out in the last section our queries are given in terms of documents and we are interested in retrieving all documents that are "close" to the given document set. To this end it is very useful to have a distance measure defined over the collection of documents and derivable from the document coupling that are based on indexing terms.

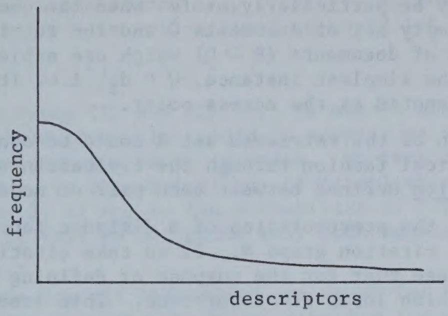


Figure 1. Descriptor Frequency Distribution

Let us consider the document descriptor matrix A which has m rows and n columns. With each row of A is associated a document and each column a descriptor. The entry a_{ij} takes the value one if the jth descriptor is assigned to the ith document and zero otherwise. We define the mxm correlation matrix as

$$C = AA^T = \|c_{ij}\| .$$

The correlation graph is defined by the following process. We assign each document a node and assign the value c_{ij} as the weight of the link between nodes i and j. Thus the weight c_{ij} of the link in the correlation graph serves as a measure of "closeness" between documents i and j.

It is noted that the number of rows of C is equal to the number of documents, m, in the document collection. This is usually a large number. To compute AA^T in the conventional way of matrix compulation would not be an attractive approach. Since the number of descriptors assigned for each individual document is small the density of entries a_{ij} in A is very low. The computation of $C = AA^T$ can then be done efficiently by list processing techniques. A detail discussion of the technique will be given in conjunction with the analysis of the citation graph in the next section.

III. THE CITATION GRAPH

Another class of structural organizations of a given collection of documents can be obtained by exploiting the bibliographic couplings. Several types of bibliographic couplings may be envisaged, such as those based on the number of shared references, citation, weighted citation, etc. Obviously, the simplest type of coupling is provided by direct citation, which may be considered as a first order association of documents. In this scheme, with each document we associate a set of documents, i.e. the documents it cites. Citation is interpreted as a directed relation between citing and cited: if we represent documents with nodes, citation can be adequately represented by directed edges from the citing document to the cited documents. We perform this representation for each document in the collection and the citation graph is constructed.

Formally, given a document collection $B \equiv \{d_1, d_2, \ldots, d_n\}$ consisting of documents $d_1, d_2, \ldots, d_n$, the directed citation graph $\mathcal{D}$ pertaining to B is entirely described by an nxn matric $E \equiv \|e_{ij}\|$, where $e_{ij} > 0$ if and only if document d_i cites document d_j.

As noted, citation indicates an association between documents and could be conveniently exploited in retrieval operations. Specifically, the citation structure may be particularly useful when the query is formulated by specifying a non-empty set of documents Q and the retrieval goal is the extraction of a set R of documents $(R \supset Q)$ which are subject-related to the documents of Q. In the simplest instance, $Q \equiv d_i$, i.e. it contains a single document d_i. d_i is denoted as the access point.

The determination of the retrieved set R could be conveniently performed in a mechanical fashion through the evaluation of some single-valued distance function defined between each pair of nodes of the graph.

Before analyzing the prerequisites of a distance function, we reconsider the directed citation graph $\mathcal{D}$. If we take citation as a sign of subject-relation, we see that for the purpose of defining subject-areas the direction of citation loses its importance. This leads us to replacing the directed graph $\mathcal{D}$ with the undirected graph $\mathcal{U}$, simply denoted as the citation graph. $\mathcal{U}$ is described by the nxn matrix

$$T = \|t_{ij}\| = C + C^T$$

where now $t_{ij} = t_{ji} > 0$ means that d_i and d_j are linked through direct citation. The weight of the linkage, t_{ij}, may be binary-valued (0,1) if we are simply interested in the presence or absence of citation. In more refined schemes it could be real-valued non-negative, its magnitude measuring the strength of coupling in a normalized interval (0,1).

We now make an attempt to formulate some properties which seem to be desirable for a distance function f_{ij} defined for every pair of nodes d_i, d_j of the graph $\mathcal{U}$: obviously f_{ij} must provide an intuitively satisfactory measure of connectivity.

First, suppose that a procedure has been given for the computation of f_{ij}. It seems reasonable to require that, if the coupling strength t_{hk} between two generic documents d_h and d_k is increased (i.e. t_{hk} is a continuous parameter), the distance between any two distinct documents d_i, d_j cannot increase. Formally, in the hypothesis that coupling strengths are continuous parameters

$$\frac{\partial f_{ij}}{\partial t_{hk}}$$

must be continuous and for $t_{hk} > 0$, $f_{ij} \geq 0$ we must have

$$\frac{\partial f_{ij}}{\partial t_{hk}} \leq 0 \qquad (1)$$

i.e. f_{ij} is a monotonically non-increasing function of the t_{hk}'s.

Secondly, assume that two documents d_i and d_j are linked exclusively through a third document d_k, i.e. that every and each path P_{ij} between d_i and d_j contains d_k. In this case it seems natural to require that the distance function f_{ij} be additive, or

$$f_{ij} = f_{ik} + f_{kj} \qquad (2)$$

We must point out, at this stage, that more than to a semantic similarity between documents we are aiming to some easily and mechanically computable correlation based on the citation association.

Returning now to our main line, we notice that the well-known function "resistance" defined over the graph $\mathcal{U}$ would meet our previous requirements (1), (2). The graph $\mathcal{U}$ is considered as a resistive network, in which each edge b_{hk} is assigned a resistance $1/t_{hk}$. Since the resistance R_{ij} between any two nodes d_i, d_j of $\mathcal{U}$ is well-defined we could let

$$f_{ij} = R_{ij} .$$

In addition to verifying (1) and (2), R_{ij} is also a metric function. Another well-known function which could be adopted as a measure of distance is the "reliability" between pairs of nodes. We recall that reliability r_{ij} between d_i and d_j is the probability of establishing a transmission path between d_i and d_j if t_{hk} is the probability of correct functioning for the edge b_{hk}. It is easy to recognize that both requirements (1) and (2) are verified by r_{ij}.

A number of topological techniques are known for the evaluation of either the resistance function or the reliability function respectively. These techniques are satisfactory for most applications. In computer based information retrieval systemshowever, the procedure must be applied many times for each retrieval operation and simplicity in methods employed is of utmost importance.

For this reason, we turn our attention to another function which can be defined for each pair of nodes of $\mathcal{U}$. We recall that a circuit is a set of m undirected edges $b_1, b_2, \ldots, b_m$, such that: i) each b_j can be oriented; ii) the terminal of b_j coincides with the origin of b_{j+1}; iii) the terminal of b_m coincides with the origin of b_1. Obviously a circuit G_{ij} containing d_i and d_j is composed of two paths which are edge-disjoint (but not necessarily node-disjoint). We can now give the following

Definition: Let $G_{ij}^{(1)}, G_{ij}^{(2)}, \ldots, G_{ij}^{(n)}$ be the totality of distinct circuits containing two distinct nodes d_i and d_j. We define as the length of the circuit $G_{ij}^{(s)}$ $(s = 1,2,\ldots,n)$

$$\ell\,[G_{ij}^{(s)}] ,$$

the sum of $1/t_{hk}$ over each edge belonging to $G_{ij}^{(k)}$. Then we let

$$f_{ij} = \min_{s} \ell\,[G_{ij}^{(s)}] . \tag{3}$$

We note that f_{ij} satisfies requirements (1) and (2). In fact if t_{hk} is the weight of edge b_{hk} and G_{ij} is a minimum length circuit, then

$$f_{ij} = \sum_{b_{hk} \in \bar{G}_{ij}} \frac{1}{t_{hk}} .$$

It follows that

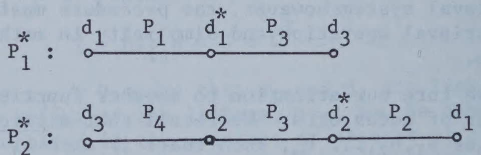
$$\frac{\partial f_{ij}}{\partial t_{hk}} = \begin{cases} 0 & \text{if } b_{hk} \notin \bar{G}_{ij} \\ -\dfrac{1}{t_{hk}^{2}} < 0 & \text{if } b_{hk} \in \bar{G}_{ij} \end{cases}$$

By letting $f_{ii} = 0$ for each i, verification of property (2) follows from the stronger statement that f_{ij}, as given by (3), is a metric function. The proof of this assertion is considerably simplified by the following lemma.

Lemma: If there is a circuit G_1 containing d_1 and d_2 and a circuit G_2 containing d_2 and d_3, then there exists a circuit containing d_1 and d_3.

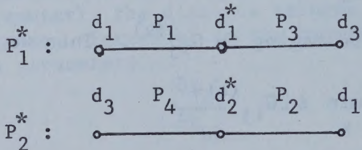
Proof: Let G_1 consist of the two edge-disjoint paths P_1, P_2 and similarly G_2 consist of P_3, P_4. Since $P_1 \cap G_2$ is non-empty, (at least they contain node d_2) starting from d_1 and proceeding on P_1, let d_1^* be the first node of P_1 which also belongs to G_2. Similarly, let d_2^* be the analogous node on P_2. We have now the following two situations:

1) d_1^*, d_2^* belong to the same path of G_2, say P_3. Then traversing P_3 from d_3 to d_2, assume, with no loss of generality, that we first reach d_1^* (if $d_1^* \equiv d_2^*$, it is immaterial which d_j^* (j = 1,2) is chosen as the first node reached). Path P_3 is therefore partitioned into paths $d_3P_3d_1^*$, $d_1^*P_3d_2^*$, $d_2^*P_3d_2$, with $d_1^*P_3d_2^*$ possibly empty. We then form the following paths P_1^*, P_2^*:



We claim that $G^* = P_1^* \cup P_2^*$ is a circuit. In fact the path $d_1P_1d_1^*$ is edge-disjoint from $d_2^*P_2d_1$ by hypothesis and from $d_3P_4d_2P_3d_2^*$ by construction (since $d_4P_1d_1^*$ contains no edge of G_2). Similarly $d_1^*P_3d_3$ is edge-disjoint from $d_3P_4d_2P_3d_2^*$ by hypothesis and from $d_2^*P_2d_1$ by construction (since the latter contains no edge of G_2).

2) d_1^*, d_2^* belong to different paths of G_2. Assume $d_1^* \in P_3$ and $d_2^* \in P_4$. Then we form the two paths



and argue as in case 1. Q.E.D.

We see therefore that f_{ij}, as given by (3), is real-valued, satisfies the reflexive property by definition and the symmetric property because of the undirectedness of $\mathcal{U}$. The triangle inequality follows from Lemma 1, since, with the same symbols, G^* consists of a subset (proper or improper) of the edges of $G_1 \cup G_2$. Hence

$$\ell[G^*] \leq \ell[G_1] + \ell[G_2]$$

and the inequality holds also when G_1 and G_2 are of minimal length. We have therefore proved

Theorem: The function f_{ij} (3) is a metric function.

In addition to some other reason which we shall mention later, an interesting feature of function (3) is the relative ease with which it can be mechanically computed.

A string S is a sequence over the set of symbols (integers) 1,2,...,n. Over the set of strings we define the operation of a string product: The string product of S_1 and S_2 is their concatenation $S_1 \cdot S_2$. Clearly, the string product is associative but not commutative. With the symbol 0 we denote the zero string, i.e. the string of no symbols. By definition, for every S, 0.S = S.0 = 0. Further a string product S is 0 in the following circumstances (nullification rules):

Rule i) S is of the form ...hk...hk... or ...hk...kh... (i.e. a given pair of consecutive symbols is repeated either in the same order or in reversed order).

Rule ii) S is of the form h...h (i.e. the first and the last symbols of S coincide).

Given these definitions, we construct the matrix M, obtained from A by replacing each $t_{hk} > 0$ with the integer k, which is now regarded as a symbol in the sense specified above.

Assume now, for simplicity, that we aim to compute the distance with respect to d_1. We multiply the first row $\underline{u}^{(1)}$ of M by M and replace the ordinary operation of multiplication with the just defined string product. We obtain the vector

$$\underline{u}^{(2)} = \underline{u}^{(1)} M .$$

We iterate this operation s-1 times and obtain

$$\underline{u}^{(s)} = \underline{u}^{(s-1)} \cdot M .$$

Let us analyze $\underline{u}^{(s)}$ for $s \geq 3$. Its first component, which is then conventionally set to 0 (rule ii), gives a collection of circuits containing d_1 and composed of s edges: in fact rules 1,2) of nullification of the string product ensure us that no edge is traversed more than once. By this iterative procedure we can obtain all circuits containing d_1 with up to s edges.

The computation of the distance becomes trivial in the particular case in which all edges are equally weighted, e.g. $t_{hk} = 1$ for any existing edge. In this case the distance is simply the number of edges of the shortest circuit containing the access node and the node under consideration. We can therefore give the following computer-oriented algorithm for the search of all documents up to distance s from a specified document where s is used as a control parameter. The algorithm takes advantage of the fact that the T matrix is in effect very sparse: while its order could be around several tens of thousands, the number of non-zero entries per row (the degree of the node) is, on the average, close to 10.

Algorithm. Each document $d_i \in B$ is specified through its accession number, for simplicity, i. With each i we associate a list L_i, i.e. a collection of integers which are the accession numbers of the documents directly linked through citation with i: the integers belonging to L_i are assumed to be naturally ordered.

Let i be the document specified by the query, i.e. the access point. With L we designate the current list: each term of L is, in general, a sum of all the string products having equal last symbol; the terms are ordered by increasing last symbol.

1. Set $r = 2$. Let $L = L_i$.

2. Let $\alpha_1, \alpha_2, \ldots, \alpha_{n_r}$ be the last symbols of the terms of L. Set $j = 1$.

3. Call from the archive list L_{α_j} and form the string product of the term ending with α_j by each term of L_{α_j}. If $j < n_r$, replace j with j+1 and repeat step 3; if $j = n_r$ go to 4).

4. Sort all string products obtained in iterations of step 3 by increasing last symbol: form new terms by adding all string products with equal last symbol. For $r > 2$, the term ending with i provides all circuits of length r.

5. Apply nullification rules i) and ii) on the list obtained in Step 4. The resulting list is the new L. If $r = s$, the algorithm terminates. If $r < s$, replace r with r+1 and return to step 2.

The described algorithm provides all circuits containing the access node and having up to s edges: the actual computation of the distance requires no further comment.

We must not overlook the possible objection that however simple the previous algorithm may appear, the length of the current list L may reach extremely high values for sufficiently high s. This geometric explosion with ratio equal to the average degree of the nodes would certainly take place if document-links were assigned at random. In our case, however, it appears that the structure of the citation network, through the strong interconnection of documents in a given subject area, acts in favor of a much milder increase: simple manual trials appear to confirm this intuition, but only more extensive experiments can have a probatory value.

Another promising feature of the circuit concept is related to the remark that possibly irrelevant documents, relatively close through citation to the access document, are excluded from the retrieved set R: the intuition, in fact, would suggest that if there is only one path from the access node to the node representative of a given document, the latter is most likely not subject-related to the query.

IV. SCHEMES FOR COMBINED RETRIEVAL

In the two previous sections we have analyzed the correlation graph and the citation graph as two structural organizations which can be conveniently exploited for document retrieval. As mentioned in the introduction, it seems very attractive to combine the power of the two structures in order to mitigate their respective shortcomings, i.e. the disturbance or "noise" caused, for example, by homographs in coordinate indexing or by careless citation.

If the query is specified by a single document (and there seems to be no conceptual difficulty in passing from single to composite queries), by following the criteria presented in Sections II and III, we can compute two distances of each document d_j from the query d_i: i.e. $f_{ij}^{(1)}$, as obtained from the correlation graph, and $f_{ij}^{(2)}$, as obtained from the citation graph. The combined distance f_{ij} must very reasonably be an increasing function of $f_{ij}^{(1)}$ and $f_{ij}^{(2)}$. The two simplest expressions of f_{ij} which we propose are

$$f_{ij} = a_1 f_{ij}^{(1)} + b_1 f_{ij}^{(2)} \quad (4)$$

$$\ln f_{ij} = a_2 \ln f_{ij}^{(1)} + b_2 \ln f_{ij}^{(2)} \quad (5)$$

where a_1, b_1, a_2, b_2 are positive constants. We remark that function (4) corresponds to the set theoretical operation of union when applied to the two graphs, while (5) corresponds to the set theoretical operation of intersection.

No insight has so far been obtained into the possible values of the constants a_1, a_2, b_1, b_2. An extensive experiment has been planned which should shed light on this aspect of the proposed scheme, as well as on further theoretical developments.

ACKNOWLEDGMENT

The authors wish to acknowledge the many stimulating discussions they have had on information retrieval with their colleagues Drs. Dewey Carroll, Sylvian Ray and Paul Weston.

USE OF ARITHMETIC CODES IN COMPUTERS

D. E. MULLER
University of Illinois
Urbana, Illinois

ABSTRACT

A procedure is given for using an error correcting code to check the arithmetic unit and main memory of a computer. Algorithms for fixed point operations are given. Problems associated with the use of floating point are treated and a method for doing arithmetic using a logarithmic representation is described.

1. INTRODUCTION

Can arithmetic codes be used effectively in computers to enhance their reliability, and if so, what is the best way to use them? In seeking to answer these questions we cannot overlook the sizable problems and difficulties associated with this application of codes. Nevertheless,our outlook is basically optimistic in that we shall concentrate on the solutions to problems rather than upon the discovery of still more reasons why codes are supposedly impractical.

While the idea of using error correcting and detecting codes in a computer is very old (it is said that much of the early work on codes was done with this application in mind), there has been little actual application. There are several reasons why this is so. Encoding, decoding, checking, and correcting all tend to be rather complicated and hence slow processes. Therefore, it is wasteful of computer time to perform such operations frequently. On the other hand, during a computation complex transformations are used when combining data. No error correction is usually possible unless checking is performed after almost every transfer or arithmetic operation. We shall see, however, that error detection with a reasonable chance of success is possible without greatly slowing the computation.

It is also true that coding is only applicable to that part of a calculation which is repetitive in some sense, or to that part of a computer which possesses some sort of regulatiry. For this reason, coding by itself, can provide only a very incomplete check on the operation of the computer.

Finally, the very fact that coding is so interesting to mathematicians may have made its use somewhat suspect. People who became interested in the mathematics of coding often found it more interesting to develop new codes and study the properties of existing ones than to try to use the existing knowledge. Conversely, computer designers have tended to be conservative and have also correctly felt that a gap exists in our knowledge between theory and practice. This gap is only now being filled.

2. CHOICE OF A CODING SYSTEM

Sometimes it is possible to use error correcting and detecting codes in a computer even though the computer was not designed with such use in mind. using a sufficiently general definition of coding, one might say that any checking scheme is a form of coding, provided it takes advantage of regularities to avoid 100% or greater redundancy in time or equipment. Thus, a simple sum check may be considered a code, but duplication of a calculation is not. Neither should we regard triplicated logic with vote taking as a method of coding.

From the point of view of equipment, the regular parts of a computer are the memory, the input and output, the arithmetic unit, and a few parts of the control: mainly those which do some sort of numerical operation. Our interest here will be centered on the arithmetic unit and the memory. Input-output coding is fairly conventional and commonly used. The use of codes where it is possible to do so in the control we shall regard as a separate problem. Neither will be treated here.

The reason that the arithmetic unit and memory are candidates for the use of codes is that the different bits in a word are handled in an essentially similar fashion. Our plan is to encode each word representing a number using redundant bits for checking purposes. Thus, only some of the many possible words will be used to represent numbers, and failures may be detected because they will usually result in words which are not valid code words.

Arithmetic codes have received some treatment in the literature, 1,2,4,7 and some algorithms for their use have been invented.[1] A few computers have even been designed which use such codes.[3] Yet, considering the importance of the subject, the publications seem to me to have been rather scant.

Arithmetic codes have the outstanding and unusual characteristic that the coding scheme remains valid even when coded numbers are added and subtracted. By changing the procedures slightly, other arithmetic operations can be designed to produce coded numbers.

Let us think of each word as representing an integer in the binary number system. It is convenient to use a system which is capable of representing both positive and negative integers uniformly, such as the two's complement or one's complement system. Valid code words are those which represent an integer of the form $N\alpha$, where N is also an integer which varies from one code word to another and may be either positive or negative, while α is an odd positive integer which characterizes the code, is called the modulus, and is a constant for all code words. Thus, we see that two code words represent two numbers $N_1\alpha$ and $N_2\alpha$, and their sum, using binary arithmetic, represents the number $(N_1+N_2)\alpha$ and is also a code word. The same property of invariance holds for such operations as subtraction, left shift without overflow, and right shift without roundoff.

The idea of arithmetic coding is thus very simple indeed. A precise description of the coding scheme requires that we make a choice of the modulus α and specify the interpretation of each code word as a number. The most natural, and in some ways the most convenient interpretation is the following. If a code word represents $N\alpha$ in the binary system, then we use it as our method of representing the integer N in the computer, or the fraction $N\alpha\ 2^{-(n-1)}$, where n is the word length. We shall call this system the integrated numerical interpretation, since the entire word is used in the representation of the number, and specific digits are not singled out as check digits. Various alternative methods have been proposed, and those which separate information digits from check digits are called check symbol methods. Other variations of the method described here can be obtained by changing the way that $N\alpha$ is represented in binary. However, all arithmetic codes have in common the fact that they use an odd modulus α, and that their code words are specified by an integer N. Some theoretical results have been obtained which tend to indicate that all codes which possess invariant properties useful in arithmetic work must be of this general type, a precise statement of this idea could probably not be set down.

Our next problem in selection of an arithmetic code concerns the choice of the modulus α. That α must be odd (at least in binary computers) is clear, since to take α as even would simply mean that our least significant bit would remain unused. While 2^n words can be written using n bits, we see that approximately $2^n/\alpha$ of these are code words. Hence, by increasing α we reduce the information content of the code, but we may also increase its error detecting and correcting capability. Just what sort of error correction is possible with arithmetic coding is an interesting question. Binary adders of conventional design may fail in such a way as to affect more than one bit position in the result, because a failure may have an effect upon the way a carry is propagated. Nevertheless, all single failures either increase or decrease the number represented by the binary interpretation of the result by a power of two. Some unconventional adders, such as those which use special circuits to speed carry propagation, do not possess this property, but we shall not be overly concerned with this restriction.

We let $N\alpha$ stand for a correctly encoded word as before. If a single failure occurs in a process which would otherwise produce a correctly encoded word, then the result will be of the form $N\alpha\underline{+}2^i$, where either sign may occur, and i is one of the integers 0,1,...,n-1. To uniquely correct such a result, there must be no relation of the form $N\alpha\underline{+}2^i = M\alpha$ and no relation of the form $N\alpha\underline{+}2^i = M\alpha\underline{+}2^j$ satisfied for any other correctly encoded word $M\alpha$, any combination of signs, and any j in the range 0,1,...,n-1. Since α is odd, the first relation cannot ever hold. This proves incidentally that all single errors are detected when α is odd and greater than 1. To satisfy the second relation, we require that $2^i\underline{+}2^j$ be divisible by α. Thus, to obtain single error correction, we simply require that no number of the form $2^k\underline{+}1$ be divisible by α when k has the range 1,2,...,n-1.

It is a well known property of odd numbers α that for each such number there is a minimum positive number β such that one of the two numbers $2^\beta+1$ or $2^\beta-1$ is divisible by α. Also, it may be shown that β is never greater than $\frac{1}{2}(\alpha-1)$ and reaches this upper bound only for certain prime α's. See the table at the end of this article. We thus see that an error correcting code may be obtained by choosing α so that the corresponding β is at least as large as n.

In defining β we said that one of the two numbers $2^\beta+1$ and $2^\beta-1$ should be divisible by α. Clearly both cannot, and we see therefore that our codes are of two types. If the number $2^\beta+1$ is divisible by α we shall call the code an odd code, and if the number $2^\beta-1$ is divisible by α we shall call it an even code. These two types of code have slightly different arithmetic properties. Using a symmetric range for N about zero, we have a maximum value of N at $\frac{1}{2}(2^\beta+1-\alpha)/\alpha$ for odd codes and at $\frac{1}{2}(2^\beta-1-\alpha)/\alpha$ for even codes. The realization of this range requires that we take $n=\beta$.

Returning to the problem of choosing α, we recognize the fact that many considerations may influence our choice of α. Some values of α which are easiest to implement turn out to be the poorest from the point of view of error correction and detection. However, for what we propose here the circuitry requirements are not too onerous, and therefore we imagine the selection of an α which is capable of providing single error correction. It may be argued that if multiple errors may occur in our result before checking, such a code is also more likely to detect errors. It will also be most convenient and useful for diagnostic purposes.

To get a maximum range for N, we choose α so as to maximize the maximum value of N indicated above, subject to the requirement that n be no greater that β in order to get at least single error correction. In the useful range for n, this usually means that α is one of those primes for which $\beta = \frac{1}{2}(\alpha-1)$. There are a few exceptions, however. For example, for a 55 bit word, it is most efficient to choose $\alpha = 121$, which is not prime. The best prime α which will work for $n = 55$ is 131, and it is capable of correcting a 65 bit word.

As might be assumed, the properties of arithmetic codes are often consequences of elementary number theory. We note the following results which are easily proved: Whenever α is of the form $8m-1$ for some integer m the corresponding code is even. Whenever $\beta = \frac{1}{2}(\alpha-1)$, and α is of the form $8m\pm3$, then the corresponding code is odd. Whenever $\beta = \frac{1}{2}(\alpha-1)$, then α is a prime, but cannot be of the form $8m+1$.[5]

3. FIXED POINT ARITHMETIC

Let us imagine that α has been chosen, and we now wish to design an arithmetic unit which allows us to perform arithmetic with code words. It was noted that the operations of addition and subtraction do preserve coding. However, this is only true if overflow does not occur, since an overflowed word is no longer correctly encoded. In ordinary arithmetic calculations, this fact should prove no great inconvenience. One case does exist in which an overflowed word remains correctly encoded. If we use an even code and take $n=\beta$, and use a one's complement representation for $N\alpha$, then the end around carry which occurs on overflow is exactly what is needed to restore the code word. To see this, we note that 2^n-1 is divisible by α. Overflow has the effect of decreasing the result by 2^n, and the end around carry increases the result by 1. The result is still divisible by α and hence correctly encoded. A similar argument shows that underflow also does not destroy coding.

While left shift without overflow may be carried out in the standard way, we must alter the process for right shifting with roundoff so that coding is preserved. An understanding of the right shift process is basic to an understanding of the more complex processes of multiplication and division. Consider the case in which $N\alpha$ is a positive odd number to be right shifted one place. A standard right shift replaces $N\alpha$ by $\frac{1}{2}(N\alpha-1)$. Since this number is not divisible by α, we must alter it in some way. This can be done by either adding $\frac{1}{2}(\alpha+1)$ to or subtracting $\frac{1}{2}(\alpha-1)$ from the shifted result. The same roundoff error is produced by either of these operations, but the errors are in opposite directions.

Our choice of a rule for right shifting is based on considerations of economy and speed. Let c be the minimum exponent such that 2^c is greater than or equal to $\frac{1}{2}(\alpha+1)$. While performing a standard right shift, we check to see whether the digit in the result of weight 2^c is 0 or 1. In the former case we add $\frac{1}{2}(\alpha+1)$ and in the latter case we subtract $\frac{1}{2}(\alpha-1)$. The rightmost $c+1$ digits are the only ones involved in the correction with a resultant improvement in speed and economy over other rules. It is also important to note that roundoff tends to be statistically unbiased, assuming equal probability for the digit of weight 2^c to be 0 or 1. Such a property is of great value in numerical work.

In case N is negative, the same rule as for the positive case will work if $N\alpha$ is expressed in the two's complement system. That this is so follows from the fact that when $N\alpha$ is odd, i.e. the least significant digit is 1, then the result of a standard right shiff is $\frac{1}{2}(N\alpha-1)$. Let us consider the one's complement system, however. When $N\alpha$ is odd, the least significant digit is 0, and the result of a standard right shift is $\frac{1}{2}(N\alpha+1)$. Thus, we must either add $\frac{1}{2}(\alpha-1)$ or subtract $\frac{1}{2}(\alpha+1)$. The same basis for choosing between addition and subtraction may be used as in the previous case. It is true, however, that this difference in the rule for positive and negative N will require more equipment than was required by two's complement right shifting.

We may now pass to a treatment of multiplication. It is clearly necessary to form correctly encoded versions of both the most and least significant halves of a product. Using our previous notation, suppose that $M\alpha$ and $N\alpha$ are two words to be multiplied and that $P\alpha$ and $Q\alpha$ are the most and least significant parts of the product respectively. We shall choose our algorithm so that $MN = P2^{n-c-1} + Q$. Begin by considering a single right shift of the partial product, represented by $P_1\alpha$ and $Q_1\alpha$. We use the previously described method for shifting $P_1\alpha$. If $P_1\alpha$ is even, then a standard right shift is used for both $P_1\alpha$ and $Q_1\alpha$ since we always have $Q_1\alpha$ an even number until the multiplication is complete. If $P_1\alpha$ is odd then the result of shifting it right is either $\frac{1}{2}(P_1+1)\alpha$ or $\frac{1}{2}(P_1-1)\alpha$. In the former case, we substract $\alpha 2^{n-c-2}$ from the result $\frac{1}{2}Q_1\alpha$ which we get from shifting $Q_1\alpha$ to the right. In the latter case, we add $c2^{n-c-2}$ to $\frac{1}{2}Q_1\alpha$. The shifted partial product is represented by a pair $P_2\alpha$ and $Q_2\alpha$, where we either have $P_2 = \frac{1}{2}(P_1+1)$ and $Q_2 = \frac{1}{2}(-2^{n-c-1} + Q_1)$ or else $P_2 = \frac{1}{2}(P_1-1)$ and $Q_2 = \frac{1}{2}(2^{n-c-1} + Q_1)$. In either case, $P_2 2^{n-c-1} + Q_2 = \frac{1}{2}(P_1 2^{n-c-1} + Q_1)$, and the shift has had the desired effect.

During the time that the partial product is being shifted right the multiplier must also be shifted right. The same register is often used for both the multiplier and the least significant part of the partial product. Strictly speaking, this may not be done for the present algorithm. However, the most significant n-c-1 bits of the multiplier register may be used as the least significant n-c-1 bits of the least significant part of the product register. Also, as we have seen, a special device for either adding or subtracting α must be attached to the leftmost c+2 bits of this latter register.

When, during multiplication, the rightmost digit of the multiplier register is 0, the right shift that we have described is performed. If this digit is 1, then it is necessary to either add $\frac{1}{2}(\alpha+1)$ to or subtract $\frac{1}{2}(\alpha-1)$ from the shifted multiplier, using the previous method for determining which operation to perform. If $\frac{1}{2}(\alpha+1)$ is added to the multiplier, then we also subtract the multiplicand from the most significant part of the partial product before shifting. If $\frac{1}{2}(\alpha-1)$ is subtracted from the multiplier we add rather than subtract the multiplicand. There are n-c-2 steps in the multiplication, followed by a special last step in which the rule for roundoff of the multiplier is reversed with respect to the digit of weight 2^c.

The multiplier has been lost during the process of multiplication, but the c+1 least significant bits of the multiplier register should be zero if the multiplier was correctly encoded to start with, and if no failure occurred during its shifting. We can thus easily test the residue of the multiplier to be certain it is zero. Other tests need not be made at this time because the coding of the multiplicand affects the correct coding of the product in its most significant half. If a special instruction is to be used to check numbers for correct coding then convenient equipment for performing this task is already present in the multiplier register.

A process for performing division will not be described here because of limited space. However, the reader should have little difficulty imagining such a process based on the same principles which were used in forming the multiplication process. It is important to note that while the remainder is often discarded after division, it now contains important information which provides us with a check that the division was performed correctly. Thus, in a checked program this information should be saved by adding the remainder to a register to be checked at some later time.

4. GENERAL CONSIDERATIONS

Although we assume that our code is capable of correcting only single errors, we are very likely to encounter multiple errors if a fault develops in the computer. There are two reasons for this. The faulty component, when used repeatedly is likely to fail more than once and therefore introduce more than one error in the calculation. Furthermore, even if only one failure occurs, later calculation is likely to proliferate the single error produced by this failure, because a single number, such as a multiplicand, is used repeatedly. In order to be able to correct all errors one would have to check each elementary operation of the computer. Such a check would be more time consuming than the elementary operation itself, would require more equipment, and would be likely to introduce more reliability problems than it solved.

If we are willing to settle for error detection, the picture is brighter, but detection will not be certain since multiple errors may appear which are not detectable by the code. However, if we make the assumption that all multiple errors are equally likely, then the probability of an undetected error is about $1/\alpha$. One might expect any bias toward single errors to reduce this probability. In fact, there is probably ample justification for using a code which is capable of correcting single errors, both from the standpoint of detection, and in order to simplify fault diagnosis. Avizienis[2] has studied the detection capabilities of various arithmetic codes.

In the model system under discussion, we shall imagine that the only automatic checking which is done is of the multiplier during multiplication, and possibly of the remainder after division. Otherwise, checking is carried out under the control of the program, but a cheap convenience would be a special check and retain residue instruction using the multiplier register. Another device which is important in detection and diagnosis is an overflow detector attached to the accumulator. Some failures may result in unexpected overflow, and a unique identification of all single errors cannot always be made unless it is known whether or not overflow has occurred. This is the reason that some odd codes appear to exceed the information theoretical limit of efficiency. For example, there are $(2^\beta+1)/\alpha$ code words, 2^β errors may be corrected, and in some cases $\beta = \frac{1}{2}(\alpha-1)$. Since $(2\beta+1)(2^\beta+1)/\alpha$ exceeds the number of words we can represent with β bits, we have an apparent contradiction. This only means, however, that correction of single errors requires knowledge of overflow, thus introducing additional check information into the system.

The same code which is used in the arithmetic unit should be used in the main memory to avoid the delay of recoding. The price for having this additional speed and reliability is about one bit per word in the memory assuming optimum coding schemes. Thus, when checking is done it may be done for the memory and arithmetic unit at once.

While the memory is thus checked in the sense that alteration of a word in the memory is usually detectable, we do not check to see that the correct word was addressed in either the case of writing or reading. A special scheme may be devised as follows for checking memory addressing. An instruction which writes a word into the memory is made to add a quantity A to this word before writing the sum. The quantity A need not be a correctly coded word but should depend upon the location into which the word is being written. When this same word is read from the memory the quantity A is automatically subtracted, giving us the correct word provided addressing was carried out correctly. We may take A as the actual address if we like, but if an error occurs during writing, but not reading, then we run the risk of using some previous result instead of the one we want, and the failure will not be detectable. Other addressing errors will be detectable. The best solution would seem to be to take A as the symbolic address to which is added the contents of the index register, if any. In this case, all addressing failures seem to be detectable. Again our check is made in the same way for this type of failure as for arithmetic and memory failures.

5. FLOATING POINT ARITHMETIC AND LOGARITHMIC ARITHMETIC

Most modern computers use built-in floating point arithmetic. Although the mechanics for using fixed point arithmetic may be extended to floating point with coded fractional part, the exponent cannot be included in the code. There are four ways in which we may deal with this problem. We may leave the exponent unchecked and accept the consequent loss in reliability. We may duplicate the exponent and its associated equipment and compare the outputs of the two circuits as a check. We may use a separate arithmetic code for the exponent part of each number. Or finally, we may represent each number by the binary representation of its logarithm and encode that arithmetically.

It is our opinion that some sort of checking should be present whenever at all feasible. Even parts of the control which possess no regular structure can be checked by massive redundancy (duplication) techniques. Therefore, if nonlogarithmic floating point is to be used, we may choose either massive redundancy or arithmetic coding. Our somewhat heretical feeling is that arithmetic coding is cheaper and simpler in practice if not conceptually. For example, if we use a total of 14 bits for our exponent, we may take $\alpha = 29$, which gives us a range of exponent from -282 to +282. The redundancy for the exponent and associated equipment should be roughly 60%.

The rather far-out possibility of using a logarithmic representation of numbers has the significant advantage of uniformity. The entire number with the exception of its sign can be included in a single residue check. Red tape and bookkeeping operations may use the nonlogarithmic representation since scaling problems are not involved. There are numerous other advantages to logarithmic representation of numbers. These include ease of multiplication, division, squaring, taking the square root, etc. None of these supposed advantages is of much importance because the processes of addition and subtraction are made so slow and difficult. Thus, a logarithmic computer would be impractical unless some method is found for solving the problem of addition and subtraction.

Let us imagine that a binary, logarithmic representation is used for numbers A, B etc. in which each number has a sign and n bits. Thus, we can form such quantities as AB and A/B very easily since ordinary addition and subtraction circuitry may be used. Let n be a small even number and let us suppose that any number A may be separated into right and left halves A_1 and A_2, each consisting of $\frac{1}{2}n$ bits. We note that the numerical product of A_1 and A_2 is just A. Suppose that we are also provided with three, fast, read-only memories, each containing $2^{\frac{1}{2}n}$ words. They consist of the following tables.

(1) If X is a number greater than or equal to 1 which is accurately repre-by its leftmost $\frac{1}{2}n$ bits, the remaining bits being zero, then these bits refer to a word in the first table which represents $1+X^{-1}$.

(2) The second table is referred to in the same way as the first table, but the contents represent $1-X^{-1}$.

(3) If Y is a number which is accurately represented by its rightmost $\frac{1}{2}n$ bits, the bits to the left of these being all 0, then the rightmost $\frac{1}{2}n$ bits refer to a word in the third table which represents 1-Y.

We now develop an algorithm for generation of A+B or A-B using this equipment. We first consider the algorithm for A+B. Assume that B is no larger than A.

(1) Form $R = A/B$ and separate R into the two parts R_1 and R_2. If $R \geq 2^n/\alpha$, then use A as an approximation to A+B. Let us define $B' = R_1/A$. Then, since $R_1R_2 = R$, we see that $R_2 = B'/B$, and $R_1 = A/B'$.

(2) Construct $(1+R_1^{-1})$, using table 1 and multiply by A giving $A(1+R_1^{-1})=A+B'$.

(3) Concurrently with step 2, use table 3 to form $1-R_2$ which we multiply by B, giving $B(1-R_2) = B-B'$.

(4) Divide the result of step 2 by step 3, giving a result $T = (A+B')/(B-B')$. If $T \geqslant 2^n/\alpha$, we use the result of step 2 as our answer. Otherwise, we go on to step 5.

(5) Form the first part, T_1 of T and use table 1 a second time to obtain the quantity $1+T_1^{-1}$ which we use as an approximation to $1+T^{-1}$. We multiply $(1+T_1^{-1})$ by the result of step 2, giving $(A+B')(1+T_1^{-1})$ which is approximately $(A+B')(1+T^{-1}) = (A+B')(1+(B-B')/(A+B')) = (A+B)$.

An identical algorithm may be used for forming A-B, except that table 2 is used in place of table 1 throughout the process. The reader may check to see that the appropriate identity holds. An analysis of the error resulting from the approximation in step 5 shows that this error is no greater than is justified by the accuracy of our representation. The error analysis will not be included here.

The time taken by the algorithm described above is assumed to be mainly due to the two memory access times required to look up the functions in the tables. Memories which cannot be written into are becoming cheaper and faster. Thus, the method may be practical for computers requiring only a short word length.

BIBLIOGRAPHY

1. Avizienis, A., *A Set of Algorithms for a Diagonosable Arithmetic Unit*. Technical Report No. 32-546, Jet Propulsion Laboratory, Pasadena, California (1964).
2. Avizienis, A., *A Study of the Effectiveness of Fault-Detecting Codes for Binary Arithmetic*. Technical Report No. 32-711, Jet Propulsion Laboratory, Pasadena, California (1965).
3. Avizienis, A., *A Diagnosable Arithmetic Processor*. Presented at the IEEE Computer Group Workshop on the Organization of Reliable Automata, February 2-4, 1966; Los Angeles, California.
4. Chien, R. T., *On Linear Residue Codes for Burst-Error Correction*. IEEE Transactions on Information Theory, Vol. IT-10 No. 2 pp. 127-133 (1964).
5. Kautz, W. H. and Elspas, B., *Single-Error-Correcting Codes for Constant-Weight Data Words*. IEEE Transactions on Information Theory, Vol. IT-11 No. 1 (1965).
6. Muller, D. E., *Evaluation of Logical and Organizational Methods for Improving the Reliability and Availability of a Computer*. Presented at the IEEE Computer Group Workshop on the Organization of Reliable Automata, February 2-4, 1966; Los Angeles, California.
7. Peterson, W. W., *Error Correcting Codes*. The MIT Press and John Wiley and Sons, Inc., New York, N.Y., pp. 236-244 (1961).

TABLE GIVING SOME CHARACTERISTICS OF ARITHMETIC CODES

α	β	parity	efficiency
3	1	-	1
5	2	-	1
7	3	+	1
9	3	-	0.75
11	5	-	1
13	6	-	1
15	4	+	0.57
17	4	-	0.5
19	9	-	1
21	6	+	0.6
23	11	+	1
25	10	-	0.83
27	9	-	0.69
29	14	-	1
31	5	+	0.33
33	5	-	0.31
35	12	+	0.71
37	18	-	1
39	12	+	0.63
41	10	-	0.5
43	7	-	0.33
45	12	+	0.55
47	23	+	1
49	21	+	0.88
51	8	+	0.32
53	26	-	1
55	20	+	0.74
57	9	-	0.32
59	29	-	1
61	30	-	1
63	6	+	0.19
65	6	-	0.19
67	33	-	1
69	22	+	0.65
71	35	+	1
73	9	+	0.25
75	20	+	0.54
77	30	+	0.79
79	39	+	1
81	27	-	0.67
83	41	-	1
85	8	+	0.19
87	28	+	0.65
89	11	+	0.25
91	12	+	0.27
93	10	+	0.22
95	36	+	0.77
97	24	-	0.5
99	15	-	0.31
101	50	-	1
103	51	+	1
105	12	-	0.23
107	53	-	1
109	18	-	0.33
111	36	+	0.66
113	14	-	0.25
115	44	+	0.77
117	12	+	0.21
119	24	+	0.41
121	55	-	0.92
123	20	+	0.33
125	50	-	0.81
127	7	+	0.11
129	7	-	0.11
131	65	-	1
133	18	+	0.27
135	36	+	0.54
137	34	-	0.5
139	69	-	1
141	46	+	0.66
143	60	+	0.85
145	28	+	0.38
147	42	+	0.58
149	74	-	1
151	15	+	0.2
153	24	+	0.32
155	20	+	0.26
157	26	+	0.33
159	52	+	0.66
161	33	+	0.41
163	81	-	1
165	20	+	0.24
167	83	+	1
169	78	-	0.93
171	9	-	0.11
173	86	-	1
175	60	+	0.69
177	29	-	0.33
179	89	-	1
181	90	-	1
183	60	+	0.66
185	18	-	0.20
187	80	+	0.86
189	18	+	0.19
191	95	+	1
193	48	-	0.5
195	12	+	0.12
197	98	-	1
199	99	+	1
201	33	-	0.33

Key to notation: Even parity is indicated by "+" and odd parity by "-".
Efficiency is defined as $2\beta/(\alpha-1)$.

DEVELOPMENT AND USE OF SENSITIVITY MODELS

P. Kokotović and J. Medanić
M. Pupin Inst. for Automation
and Telecommunications
Belgrade, Yugoslavia

R. Rutman
Institute for Machines Theory
USSR Academy of Sciences
Moscow, USSR

Introduction

The sensitivity problem is expressed usually in terms of the influence of a variation in a given part of the system on the behavior of the whole system. In the simplest and the most widely studied case a time invariant parameter is considered as a part of the system, and the influence of a small change of its numberical value is analyzed [1].

One can expect that in a less general class of systems, a more general type of variation may be analyzed, and vice versa [2]. In linear systems, for example, the effects of bounded but otherwise arbitrary variations have been analyzed and some closed formulas for signal and parameter invariancy or controller compatibility have been obtained [3,4,5]. However, if the system is described by a set of nonlinear operator relations, the only reasonable approach is, evidently, the use of the continuity and differentiability properties of some nonlinear operators and differential equations, and, consequently, the assumption that the variations are small or, at least that they may be expanded in some series.

In the case of small variations it is still important to distinguish parametric (i.e. numerical, time-invariant) and functional variations. Although the first is a special case of the second one, the practical applicability of the two may differ considerably. The simplest example is a circuit containing a nonlinear element, for example, a diode, designed under the assumption that the diode is ideal. Let the task of sensitivity analysis be to estimate the effects produced by the departure of the diode characteristic from the ideal. If this departure may be expressed only by the presence of a non-zero direct resistance and a non-zero inverse conductivity, the problem is reduced to a parametric variation. If one is interested in the more realistic form of distortion, the variation is of a more complex type and may not be expressible by the change of numerical values of some parameters.

Continuing the discussion concerning small variations, we may ask for what purposes the knowledge of their effects is to be used. In other words, is sensitivity (or better, insensitivity) the goal of system synthesis, or is the sensitivity method a tool for some other design purposes? Surveying existing results from the field of sensitivity [1], the answer might be - both. In the design of optimum systems, for example, various sensitivity functions and indices serve for further reduction of system sensitivity or for a direct synthesis of insensitive systems [6,7]. On the other hand, the knowledge of the effects of small variations indicates how to choose the system structure [8-11] and how to alter parameters to improve performance.

The sensitivity problem leads to an operation which means, in a sense, the differentiation of the whole system operator with respect to the operator of a system element or a subsystem. From this point of view, a parallel between modern system theory and classical mathematical analysis can be notices: sensitivity analysis seems to take the place in system theory which is taken by differential calculus in mathematical analyses. To demonstrate how a variational approach unifies different problems in system theory, we can consider a performance index $J(u)$, which is a functional of a function u. Let the variation δJ correspond to the variation δu. Depending on the role of δu, the following conditions on δJ may be imposed.

PROBLEM	CONDITION	ROLE OF δu
Optimality, min. max.	$\delta J \geqq 0$ $\delta J \leqq 0$	control
Invariancy	$\delta J \equiv 0$	disturbance
Insensitivity	$\delta J \equiv 0$	parameter
Controllability	$\delta J \not\equiv 0$	control

The main role of sensitivity models is to make apparent the structure of the variations of performance indices and system variables, and to serve as a tool for the analytical, numerical, experimental and simulational methods for analysis and optimization. From this point of view some properties of sensitivity models are discussed in this paper.

Structural Rule

We will use a simple rule to compare the structure of the sensitivity model with the system structure. The rule consists in the formation of the sensitivity models for each system element separately, and then in connecting them as the corresponding elements are connected in the basic system [12,15].

Let $z \in Z$ and $y \in Y$ be the input and the output of a system element, and let the dependence of the output y on the input z and a parameter q, $q \in Q$, be defined implicitly by the relation

$$R(z,y,q) = 0 \ , \qquad (1)$$

where R is a nonlinear operator (see Figure 1).

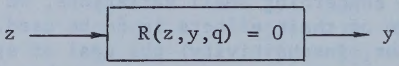


Figure 1

For instance, R may represent a system of differential equations and y its solution. We assume that a specific (nominal) solution y_o, corresponding to a particular (nominal) pair z_o, q_o, can be found either experimentally or numerically. This does not mean, however, that we know the analytic expression for the operation H, $y = H(z,q)$. Except in trivial cases, the operation H is never known explicitly. Nevertheless, having y_o, z_o, and q_o and the implicit relation (1), we are able to find the relation for the small variation δy caused by the simultaneous small variations δz and δq. We assume that R is Frechet differentiable with respect to y, z, and q, and we denote corresponding strong partial derivatives by R_y, R_z and R_q. Assuming also that the inverse R_y^{-1} exists, we will have, on the basis of implicit function theorem:

$$\delta y = - R_y^{-1} (R_z \ \delta z + R_q \ \delta q) \ , \tag{2}$$

where δy, δz, and δq are small elements of Y, Z and Q, respectively. The expression (2) may be represented by the block diagram, Figure 2. In the sensitivity model we will consider the dotted block in Figure 2 as corresponding to the block R of the system. The quantity δz plays the role of the input z, and δy is the output analogous to y. Depending on the system configuration, the output y of the system element R may be the input to some other system elements. Consequently, the output δy will be the input to the analogous parts of the sensitivity model. The similar conclusion applied to δz and z viewed as outputs. The sensitivity model of a feedback system (Figure 3a) will illustrate the rule. It is assumed that only the plant P contains the parameter q, and that the input r to the controller C does not depend on q, i.e. $\delta r = 0$. By applying the rule of Figure 2 to C and P we obtain the structure of Figure 3b.

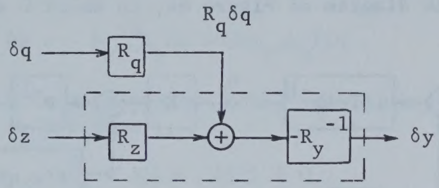


Figure 2

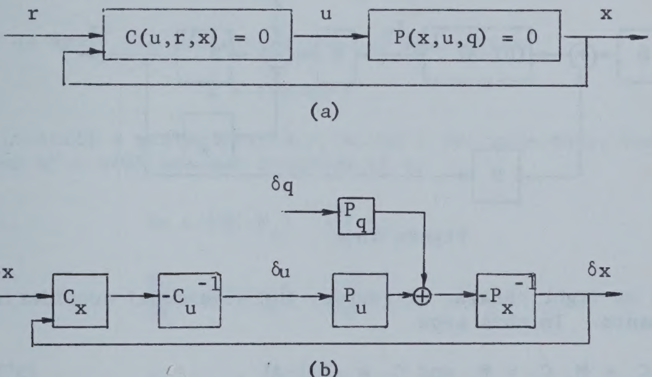


Figure 3

It follows, therefore, that, as far as the connection of elements is concerned, the sensitivity model has the same structure as the basic system. However, there are two differences. First, to the blocks containing parameter q additional input of the type $R_q\delta q$ have to be applied, and, second, all operators in the sensitivity model are linear.

Let's suppose now, that in the previous example, we are given a more detailed description of blocks C and P, say,

$$C(x,r,u) = -(DI-A)u + Mr + Bx = 0 , \tag{3}$$

$$P(x,u,q) = -Dx + f(x,q) + Nu = 0 , \tag{4}$$

where $q = q(t)$ is a scalar, $D = d/dt$, and A,B,M,N are matrices and u,x,r and f are vectors of appropriate dimensionality. The system may be represented by the block diagram of Figure 4a, in which C and P are decomposed into elementary blocks. All blocks except $f(x,q)$ are of the type $R(z,y) = Gz-y = 0$, so that $R_z = G$ and $R_y = -1$. Hence, all these blocks will correspond to identical blocks in the sensitivity model, see Figure 4b. The Jacobian matrix F_x and an input block $\partial f/\partial q$ correspond to the element $f(x,q)$.

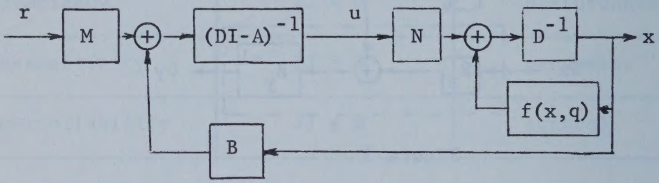


Figure 4a

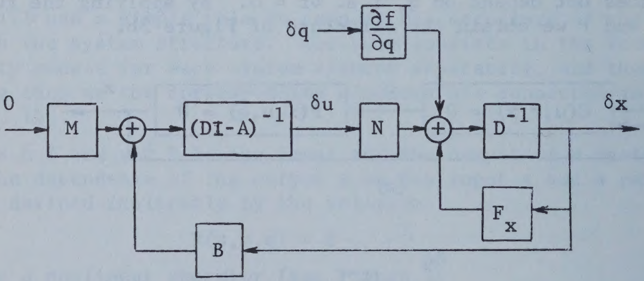


Figure 4b

In this example we might obtain, of course, the same model considering only C and P as elements. In this case

$$C_r = M,\ C_x = B, \text{ and } C_u = -(DI-A) , \tag{3'}$$

$$P_u = N,\ P_q = \frac{\partial f}{\partial q} , \text{ and } P_x = -DI + F_x , \tag{4'}$$

which is in accord with the Figure 4b. It follows, for the general case, that the whole linear part of the system remains unchanged in the sensitivity model, while instead of memoryless nonlinear elements, we have their Jacobians.

Sensitivity Operator

In a system defined by

$$S(x,r,q) = 0 \ , \tag{5}$$

where the input does not depend on parameter q, we will have, on the basis of (2):

$$\delta x = - S_x^{-1} S_q \delta q = U(x_o,q_o)\delta q \ . \tag{6}$$

The linear operator U is referred to as the sensitivity operator [2]. If δq can be expressed in the form

$$\delta q = \gamma h(t) \, , \; h(t) \in Q \ , \tag{7}$$

where $\gamma \to 0$ is a time-invariant parameter, we will have for δx

$$\delta x = \gamma \frac{\partial x}{\partial \gamma} \, , \; \frac{\partial x}{\partial \gamma} = - S_x^{-1} S_q h = Uh \ . \tag{8}$$

The weak differential $\partial x/\partial \gamma$ can be treated as a parametric sensitivity function with respect to γ, but it will depend on h(t).

Finally, if $\delta q = \Delta q$ is a small time-invariant change, the sensitivity operator coincides with the parametric sensitivity function:

$$U = \frac{\partial x}{\partial q} = - S_x^{-1} S_q \ . \tag{9}$$

Let us apply the above relations to the system

$$\dot{x} = f(x,q,t) \ , \tag{10}$$

where q is still a scalar quantity, x and f are n-vectors, and initial conditions $x^o = x(0)$ are not function of q:

$$\delta x = [DI - F_x]^{-1} \frac{\partial f}{\partial q} \delta q \ , \tag{6'}$$

$$\frac{\partial x}{\partial \gamma} = [DI - F_x]^{-1} \frac{\partial f}{\partial q} h \ , \tag{8'}$$

$$\frac{\partial x}{\partial q} = [DI - F_x]^{-1} \frac{\partial f}{\partial q} \ . \tag{9'}$$

Here, F_x is Jacobian of f with respect to x, and δx, $\partial x/\partial \gamma$, and $\partial x/\partial q$ are zero at $t = 0$.

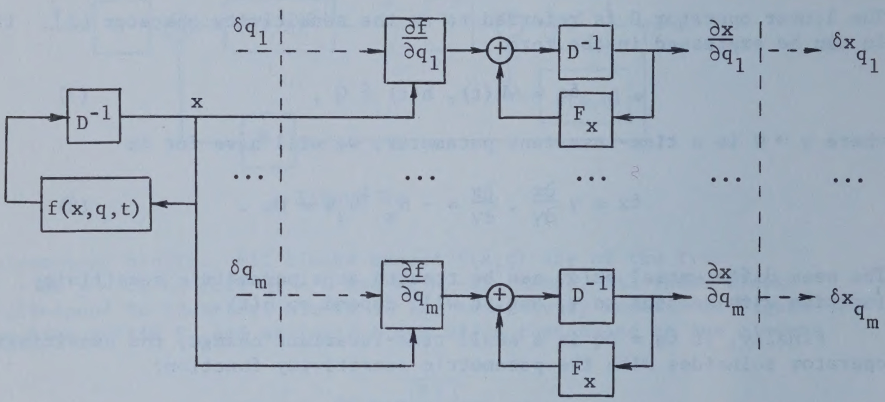
If the parameter q is a time-invariant m-vector and Δq is its parametric variation, we may write again for the first variation Δx

$$\Delta x = U \Delta q \ , \tag{11}$$

but now the sensitivity operator U is represented by an nxm sensitivity matrix. The elements of the matrix are sensitivity functions $\partial x_i/\partial q_k$, and columns are the sensitivity vectors $\partial x/\partial q_k$. The rows are the transpose of the "sensitivity gradients" of $x_i(t,q)$. For the system (10) we may then write

$$U = [DI-F_x]^{-1}\, F_q \ , \tag{12}$$

where F_q is nxm Jacobian matrix of $f(x,q,t)$ with respect to q. The block diagram for the relation (12) is given in Figure 5.



δq_1
$\frac{\partial f}{\partial q_1}$
+
D^{-1}
$\frac{\partial x}{\partial q_1}$
δx_{q_1}
D^{-1}
x
F_x
...
f(x,q,t)
δq_m
$\frac{\partial f}{\partial q_m}$
+
D^{-1}
$\frac{\partial x}{\partial q_m}$
δx_{q_m}
F_x

Figure 5

For obtaining the sensitivity matrix, experimentally or numerically, m identical sensitivity models or m computer runs are needed. The dotted lines show that the same model will provide strong variations δx, if inputs $\delta q_1, \ldots, \delta q_m$ are applied.

Special Cases

There are two important cases in which the possibility of applying the structural rule and obtaining the sensitivity matrix is not evident at first glance.

(a) Change in system order. A set of small "parasitic" parameters is always present in a physical system so that its order is higher than the order of the mathematical model. On the other hand, system designer often intentionally simplifies mathematical model by neglecting some parameters and equations. Could we obtain sensitivity functions indicating the influence of neglected parameters? Is it possible to obtain sensitivity functions from the lower order models in which these parameters are neglected?

For a class of nonlinear systems including the system

$$\dot{x} = f(x,z,t) \ , \qquad x,f = \text{n-vectors} \ , \tag{13}$$

$$\mu\dot{z} = \varphi(x,z,t) \ , \qquad z,\varphi = \text{m-vectors} \ , \tag{14}$$

where μ is a small parameter with a "nominal" value $\mu = 0$, Vasileva [14] formulated the conditions under which the functions $\partial x/\partial\mu$ exist and the quantity $\mu\ \partial x/\partial\mu$ may be used to approximate Δx everywhere, except at $t = 0$, where $\partial x/\partial\mu$ has a jump.

We have shown that in the case μ belongs to the linear part of a nonlinear system, the structural rule, Figure 2, is applicable and gives the jump conditions automatically [13]. The idea consists in representing the element containing μ by a closed loop containing μ^{-1} as a factor in the loop gain. This is the case with the element in Figure 6a, where A and B are differential operators and B is of the lower order. The function $\partial x/\partial\mu$ will exist if the loop in Figure 6a is stable for infinite gain. If this is the case, we can first apply the rule of Figure 2, and then set $\mu \to 0$. The result will be two identical lower order ("degenerate") elements, the first one in the system model and the second one in the sensitivity model, Figure 6b. It is important to note that the connection between the system and the sensitivity models is performed by the operator A, which no longer appears either in the system model or in the sensitivity model. Hence, by using two "degenerate" models we may predict what will be the behavior of the higher order system.

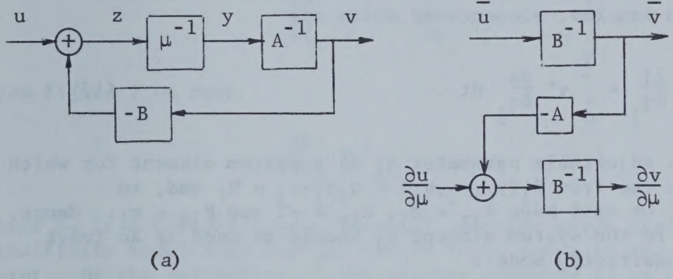

u
z
y
v
μ^{-1}
A^{-1}
-B
$\bar{u}$
$\bar{v}$
B^{-1}
-A
$\frac{\partial u}{\partial \mu}$
B^{-1}
$\frac{\partial v}{\partial \mu}$
(a)
(b)

Figure 6

(b) Discontinuous Nonlinearities. The next important case involves the presence of the discontinuities of the first kind (relays elements) in the functions $f(x,q,t)$ in (10). Even in that case, with some precaution, it is possible to proceed in the way described in this paper, but, by allowing the Jacobian matrix F_x to contain Dirac functions of x [19]. Such a sensitivity model is still realizable if the elements with impulse functions are followed by integrators.

On Simultaneity in Gradient Optimization

One application of sensitivity models is for iterative or continuous minimization of system performance indices. Let J be a functional

$$J = \int_0^T g(x)\ dt\ , \qquad (15)$$

and $g(x)$ be a concave scalar function of the system output signal n-vector x. We assume that in the controller, m parameters are adjustable so that $x = x(t,q)$, and $J = J(q)$, where q is an m-vector. The problem is to determine the vector gradJ which can be used for some convergent minimization procedures. Although this is the formulation of a parameter optimization problem, the properties of sensitivity models, discussed in this paper, indicate that an analogous reasoning is applicable for the optimization in function spaces.

Differentiating (15) we will have

$$\frac{\partial J}{\partial q} = (\text{grad} J)' = \int_0^T v' \frac{\partial x}{\partial q}\, dt\,, \tag{16}$$

where $v = \text{grad}_x g$ and the superscript denotes transposition. In this notation $\partial J/\partial q$ is an lxm row, elements of which are

$$\frac{\partial J}{\partial q_j} = \int_0^T v' \frac{\partial x}{\partial q_j}\, dt\,. \tag{17}$$

We can consider each adjustable parameter q_j as a system element for which the relation (1) has the from $R_j(y_j, z_j, q_j) = q_j z_j - y_j = 0$, and, in accordance with (2), we will have $R_{jz} = q_j$, $R_{jy} = -1$ and $R_{jq} = z_j$. Hence, the input signal z_j to the system element q_j should be used as an input signal to the jth sensitivity model.

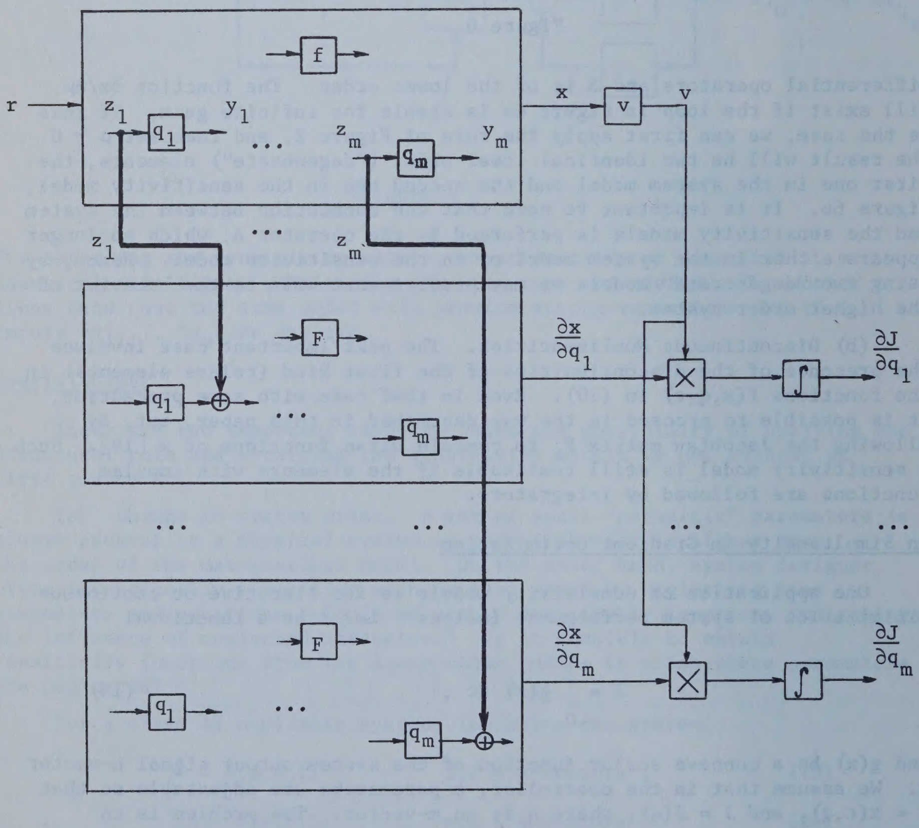

f
r
z_1
y_1
q_1
...
z_m
y_m
q_m
x
v
v
z_1
...
z_m
F
q_1
...
q_m
$\frac{\partial x}{\partial q_1}$
$\frac{\partial J}{\partial q_1}$
...
...
...
...
F
q_1
...
q_m
$\frac{\partial x}{\partial q_m}$
$\frac{\partial J}{\partial q_m}$

Figure 7

The scheme for the simultaneous generation of all components of gradJ is given in Figure 7, where f represents nonlinear elements and $F = F(t)$ represents their Jacobians. The disadvantage of this scheme is that it requires a separate model for each parameter. In a class of linear time-invariant systems the method of sensitivity points eliminates this difficulty [16]. The difficulty can be overcome in general case by using the adjoint model, as has been proposed by Narendra and McBride [17] and generalized by Medanić [18].

The shortest way to demonstrate the idea of applying the adjoint sensitivity model is to look at (17) as a scalar product in Hilbert space:

$$\frac{\partial J}{\partial q_j} = \langle v, \frac{\partial x}{\partial q_j} \rangle . \tag{18}$$

From Figure 7 we have

$$\frac{\partial x}{\partial q_j} = L_j z_j \tag{19}$$

where L_j is a bounded linear integral operator which describes the sensitivity model from the jth scalar input point to the vector output point. By the definition of the adjoint operator L_j^* we have

$$\langle v, L_j z_j \rangle = \langle L_j^* v, z_j \rangle, \text{ and,} \tag{20}$$

$$\frac{\partial J}{\partial q_j} = \langle \eta_j(t), z_j(t) \rangle, \text{ where } \eta_j(t) = L_j^* v(t) . \tag{21}$$

Now, we reverse the inputs and outputs of each block in the sensitivity model and replace the time t of any time-varying element by $T-t_1$. For example, instead of the matrix $F = F(t)$, we will have its transpose with the elements reversed in time, $F_1' = F'(T-t_1)$. If we apply the signal $v(T-t_1)$ to the input of such a model, the signals $\eta_j(T-t_1)$, $j = 1,\ldots,m$ will appear at locations corresponding to previous m input points. Considering

$$\frac{\partial J}{\partial q_j} = \langle \eta_j(T-t_1), z_j(T-t_1) \rangle , \tag{22}$$

we come to the final block-diagram for the generation of gradJ, Figure 8.

Because of the time-reversal the scheme in Figure 8 needs two computer runs, but only one sensitivity model.

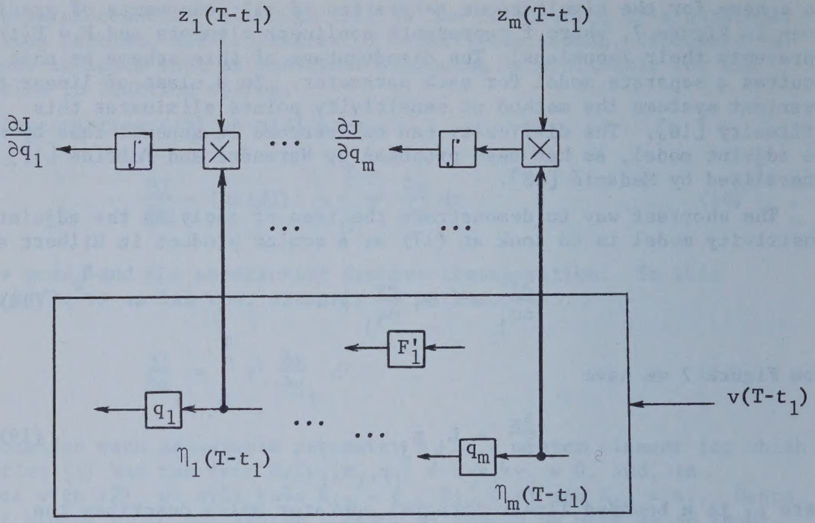


Figure 8

References

1. P. Kokotović and R. Rutman, "Sensitivity of Automatic Control Systems (survey)," Automation and Remote Control, No. 4, Vol. 26, April, 1965.
2. R. Rutman and P. Kokotović, "On the Generalization of Sensitivity Equations," Proceedings of the Third Allerton Conference on Circuit and System Theory, October 1965, University of Illinois, Urbana, Illinois.
3. L. I. Rozonoer, "A Variational Approach to the Problem of Invariance of Automatic Control Systems," Automation and Remote Control, Vol. 24, Nos. 6,7, June-July 1963.
4. R. Rutman and M. Epelman, "Parametric Invariance of Linear Systems," Dokl, Akad, Nauk SSSR, Vol. 159, No. 4, 1964.
5. J. B. Cruz, Jr. and W. R. Perkins, "On Invariance and Sensitivity," 1966 IEEE Convention Record, Part 7.
6. P. Kokotović, S. Bingulac and J. Medanić, "Some Approaches to the Reduction of Control System Sensitivity," Proceedings of the Third Allerton Conference, October 1965, University of Illinois, Urbana, Ill.
7. H. D'Angelo, M. L. Moe and T. C. Hendricks, "Trajectory Sensitivity of an Optimal Control System," Proceedings of the Fourth Allerton Conference, October 1966, University of Illinois, Urbana, Illinois.
8. W. R. Perkins and J. B. Cruz, Jr., "Sensitivity Operators for Linear Time-Varying Systems," Proceedings of the First IFAC/ETAN Sensitivity Symposium, Dubrovnik, 1964, Ed. L. Radanović, Pergamon Press.
9. P. Kokotović, J. Medanić, M. Vusković, and S. Bingulac, "Sensitivity Method in the Experimental Design of Adaptive Control System," Third IFAC Congress, London, June 1966.
10. S. Bingulac, "On the Compatibility of Adaptive Controllers," Proceedings of the Fourth Allerton Conference, October 1996, University of Illinois, Urbana, Illinois.

11. H. S. Witsenhausen, "On the Sensitivity of Optimal Control Systems," IEEE Trans., Vol. AC-10, No. 4, October 1965.

12. M. Vusković and V. Cirić, "Structural Rules for the Determination of Sensitivity Functions of Nonlinear Systems," Proceedings of the first IFAC/ETAN Sensitivity Symposium, Dubrovnik 1964, L. Radanovic, Editor, Pergamon Press.

13. P. Kokotović, R. Rutman, "On the Determination of Sensitivity Functions with Respect to the Change of System Order, Ibid.

14. A. B. Vasileva, "On the Differentation of the Solution of Differential Equations," Dokl. Akad. Nauk. SSSR, Vol. XI, No. 4, 1948 and Vol. XXV, No. 4, 1950.

15. P. Kokotović and R. Rutman, "Sensitivity Matrix and Its Simulation," Automatic and Remote Control, No. 6, June 1966.

16. P. Kokotović, "Method of Sensitivity Points in the Investigation and Optimization of Linear Control Systems," Automation and Remote Control, Vol. 25, No. 12, December 1964.

17. K. S. Narendra and L. E. McBride, "Multiparameter Self-Optimizing Systems Using Correlation Techniques," IEEE Trans., Vol. AC-9, No. 1, January 1964.

18. J. Medanić, "Application of the Adjoint System in Parameter Optimization," (in Serbo-Croatian) Pupin Institute Report, No. JM-3, P. O. Box 906, Belgrade, Yugoslavia, April 1966.

19. Ya. Z. Tsypkin, R. S. Rutman, "Sensitivity Equations for Discontinuous Systems," Automatika, No. 1, 1965, Zagreb, Yugoslavia.

SPECTRAL FACTORIZATION USING THE RICCATI EQUATION

B. L. HO
Aerospace Corp.
El Segundo, California

R. E. KALMAN
Stanford University
Stanford, California

ABSTRACT

A computational algorithm to (a) test if a polynomical $C(\omega^2)$ in ω^2 is positive semidefinite for all real ω and, if so, (b) to factor it into the product $\psi(i\omega)\psi(-i\omega)$, where $\psi(s)$ has only roots with nonpositive real parts, is developed using a new result for the existence of asymptotic optimal control of a linear regulator problem with indefinite cost functional integrand. This algorithm should be useful for checking the positive-realness of a rational function as encountered in passive network synthesis and in nonlinear stability analysis problems. It should also be useful in control and filter optimization problems.

INTRODUCTION

Many problems in circuit theory, control theory, and filtering theory lead to the problem of positive semidefiniteness of a polynomial $C(\omega^2)$, where ω are real, and the problem of factoring it into the product $\psi(i\omega)\psi(-i\omega)$, where the roots of $\psi(s)$ are to have only nonpositive real parts (the spectral factorization problem). In circuit theory, this problem arises from the realizability condition of a passive network from a driving-point impedance function [4]. In order that an impedance function $T(s) = \phi(s)/\psi(s)$ can be realizable as a passive network, it must be a positive-real function. A part of the positive-real test is to check the positive semidefiniteness of the real part of $T(i\omega)$

$$2\text{Re}T(i\omega) = \frac{\phi(i\omega)}{\psi(i\omega)} + \frac{\phi(-i\omega)}{\psi(-i\omega)} = \frac{C(\omega^2)}{\psi(i\omega)\psi(-i\omega)} \tag{1}$$

for all real ω or, equivalently, that the numerator $C(\omega^2)$ is positive semidefinite. The so-called Lur'e problem in nonlinear stability theory also calls for the positive-realness of a certain rational function [5]. Also, in control and filtering theory, it is well known that optimization in the frequency domain may be performed by the process of spectral factorization [6][8].

A method is presented here which is roughly the reverse of the last-mentioned technique. We will perform the positive semidefiniteness test and spectral factorization by making an optimization problem out of $C(\omega^2)$.

The justification of such a procedure is that powerful computer programs are now available for optimal linear control synthesis (using the matrix Riccati equation), such as the Automatic Synthesis Program described in [3], and some features of this approach are attractive as discussed in the conclusions.

There are, of course, other techniques available to perform this task -- such as finding all roots of $C(\omega^2)$ to see if all real roots in ω are in pairs. We wish to present our method simply as an additional technique to solve the stated problem and as an example of the application of modern control theory. We wish to leave the comparisons of the various techniques to the readers.

PROBLEM STATEMENT

Given the polynomical $C(\omega^2)$ with real coefficients:

(1) Is $C(\omega^2)$ positive semidefinite, $C(\omega^2) \geq 0$, for all real ω?

(2) If $C(\omega^2)$ is positive semidefinite, then we wish to factor it into the product of two polynomials with real coefficients

$$C(\omega^2) = \psi_k(i\omega)\psi_k(-i\omega) \tag{2}$$

where all roots of $\psi_k(s)$ are to have nonpositive parts.

It may be shown that problem (2) always has a solution (see, for example, [1]).

THE APPROACH

The solution is in three steps:

Step (1) - $C(\omega^2)$ is associated with the cost of a linear optimal regulator problem for a certain arbitrary (with minor restrictions) linear dynamical system.

Step (2) - The existence of an asymptotic solution of the regulator problem is equivalent to the positive semidefiniteness of $C(\omega^2)$ by Theorem 1. (discussed under "Solution to the Problem"). In computation, one periodically checks the matrix solution P of the Riccati equation associated with the optimization problem; the scalar $x_o' P x_o$ is compared with a pre-calculated constant $f(x_o)$. The scalar $x_o' P x_o$ decreases in value monotonically (with suitably chosen initial P), and if the scalar becomes smaller than $f(x_o)$ it may be concluded that $C(\omega^2)$ is not positive semidefinite. On the other hand, if $C(\omega^2)$ is indeed positive semidefinite, the scalar must remain above $f(x_o)$ and eventually reach a steady state. Then, an asymptotic optimal control law (a time-invariant control law) may be computed.

Step (3) - The closed-loop characteristic equation, $\psi_k(s)$, using the asymptotic optimal control law is the desired factorization. In computation, as mentioned in Step (2), this means that the Riccati equation is integrated long enough for the steady state solution. The speed of convergence can be controlled by properly choosing the dynamical system (system with fast transient decay for fast convergence) in Step (1).

SOLUTION OF THE PROBLEM

We will now proceed to solve the problem in detail. For those readers who are mainly interested in the method's application, the following equations are to be used in this sequence: (3, 4, 5a, 6, 8, 27, 22, 26, 32, 31, 30).

Selecting the Dynamics

Let the given polynomial be

$$C(\omega^2) = \sum_{i=o}^{n} c_i(\omega^2)^{n-i} \quad , \qquad c_o = 1 \tag{3}$$

An arbitrary characteristic polynomial is chosen,

$$\psi(s) = \prod_{i=1}^{n} (s + \lambda_i) = \sum_{i=o}^{n} a_i s^{n-i} \quad , \qquad a_o = 1 \quad . \tag{4}$$

which is stable and of degree n. In practice, λ_i are to be some large negative real numbers.

$\psi(s)$ may be considered as the characteristic polynomial of the linear dynamical system

$$\dot{x} = Fx + gu \quad , \qquad y = h'x$$
$$\psi(s) = \mathrm{Det}(sI - F) = s^n + a_1 s^{n-1} + \cdot \cdot \cdot \cdot + a_n \tag{5}$$

where x is the n x 1 state vector, u the scalar input, and F and g are n x n and n x 1 real matrices, respectively, y the scalar output, h' a 1 x n matrix.

Let us choose the following canonical forms:

$$F = \begin{bmatrix} 0 & 1 & \cdot & 0 \\ & 0 & \ddots & 1 \\ -a_n & \cdots\cdots & -a_2 & -a_1 \end{bmatrix} , \qquad g = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{5a}$$

$$h' = [h_1 h_2 \cdots\cdots h_n] \tag{5b}$$

The transfer function may be easily written from the canonical form:

$$T(s) = h'(sI - F)^{-1} g = \frac{h_n s^{n-1} + h_{n-1} s^{n-2} + \cdot \cdot \cdot + h_1}{s^n + a_1 s^n + \cdot \cdot \cdot + a_n}$$

Constructing the Cost Functional

A cost functional is constructed from $C(\omega^2)$ and $\psi(s)$ in the following way. An integral is formed, with the quotient

$$\frac{C(\omega^2)}{\psi(i\omega)\psi(-i\omega)} = 1 + \frac{q_n(-\omega^2)^{n-1} + q_{n-1}(-\omega^2)^{n-2} + \cdot \cdot \cdot + q_1}{\psi(i\omega)\psi(-i\omega)} , \tag{6}$$

as integrand, to be the cost functional plus another term:

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \left[1 + \frac{q_n(-\omega^2)^{n-1} + q_{n-1}(-\omega^2)^{n-2} + \cdot \cdot \cdot + q_1}{\psi(i\omega)\psi(-i\omega)} \right] u(i\omega)u(-i\omega)\, d\omega$$

$$= \int_{-\infty}^{\infty} [u^2 + x'Qx]\, dt = \int_{-\infty}^{0} + \int_{0}^{\infty} [u^2 + x'Qx]\, dt \tag{7}$$

where

$$Q = \mathrm{Diag}[q_1 q_2 \cdot \cdot \cdot q_n] \tag{8}$$

Q is in general not positive definite. The identity is true by the Parseval's formula provided (a) the integrals exist (this we will make certain when we formulate the optimal regulator problem), (b) that

$$x(t) = 0 \quad , \quad u(t) = 0 \quad \text{for } t < t_o < 0 \tag{9}$$

and (c) the canonical form Eq. (5a) is used going from the frequency domain to time domain in Eq. (7). Introducing the notation

$$2V(x_o, t_o, t_1; u) = \int_{t_o}^{t_1} [u' Ru + x' Qx]\, dt \quad , \quad x(t_o) = x_o \tag{10}$$

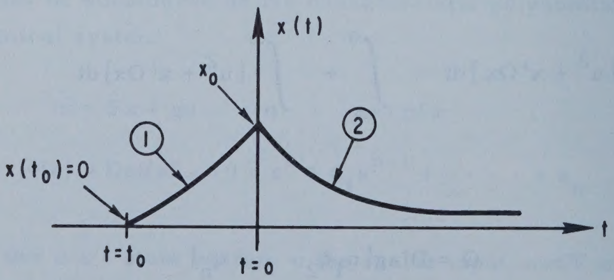
where x is the solution of Eq. (5) under the control u. We obtain the following cost functional (for the optimal regulator problem) by rewriting Eq. (7),

$$V(x_0, 0, \infty; u) = \frac{1}{4\pi}\int_{-\infty}^{\infty} \frac{C(\omega^2)}{\psi(i\omega)\psi(-i\omega)}\, u(i\omega)u(-i\omega)\, d\omega - V(0, t_o, 0; u_r) \tag{11}$$

The control u_r in the last term of Eq. (11) is any control that drives the state from 0 at $t = t_o$ to x_o at $t = 0$. That this can be done we will make sure in the optimal problem formulation by specifying that the linear system is to be controllable (a matter of choosing the right canonical form of Eq. (5a)).

The Positive Semidefiniteness Test

The meanings of the terms in Eq. (11) are clarified in the following figure



The first term $V(x_o, 0, \infty; u)$ is the cost for the curve marked 2, the third term $V(0, t_o, 0; u_r)$ for the curve marked 1, while the frequency integral is the cost for the whole curve, 1 plus 2. The important observation made from Eq. (11) is that the cost $V(x_o, 0, \infty, u)$ is bounded from below for all control u if $C(\omega^2)$ is positive semidefinite, because the cost for Curve 1, $V(0, t_o, 0; u_r)$, is a constant once u_r is chosen, and that u_r may be chosen independent of u. This assures the existence of an asymptotic optimal control as stated in the following theorem. The detail proof of the theorem essentially makes use of the above observation but is laborious -- best left for a future paper. The sufficient part of the following theorem has also been obtained independently by Popov [2].

We will denote by $\phi(t; x_o, u)$ the solution x(t) of Eq. (5) upon using u(t) starting from x_o at t_o, and $\Phi(i\omega)$ the matrix

$$\Phi(i\omega) = (i\omega I - F)^{-1} \tag{12}$$

Note that the transfer function of Eq. (5) is $h'\Phi(s)g$.

Theorem 1. If the linear dynamic system

$$\dot{x} = Fx + gu \tag{13}$$

is stable (F has eigenvalues with negative real parts), and the pair (F, g) is controllable (may be checked by their canonical forms), then the asymptotic regulator problem

$$2V^* = \lim_{T\to\infty} \min_{u} \int_0^T [Ru^2 + \phi' Q\phi(t; x_o, 0, u)]dt \quad ,$$

where R is a positive scalar and Q an indefinite matrix, with the boundary conditions

$$x(0) = x_o \quad , \quad x(T) = 0 \quad , \tag{14}$$

has a solution if and only if

$$R + g'\Phi'(i\omega)Q\Phi(i\omega)g \geq 0 \qquad \text{for all real } \omega \tag{15}$$

Note: When Q is a diagonal matrix, diag $[q_1 q_2 \cdot \cdot \cdot q_n]$, and the canonical form Eq. (5a) is chosen, Eq. (15) may be written as

$$R + g'\Phi Q \Phi g = R + \frac{q_n(-\omega^2)^{n-1} + q_{n-2}(-\omega^2)^{n-2} + \cdots + q_1}{\psi(i\omega)\psi(-i\omega)} \geq 0 \qquad (16)$$

$$\psi(i\omega) = \det \Phi^{-1}(i\omega) = \det(i\omega I - F) \quad .$$

We also have the following corollary concerning the optimal cost.

Corollary. The optimal cost of Theorem 1 is (a) bounded from below

$$V(x_o, 0, \infty; u^*) \geq f(x_o) = -V(0, t_o, 0; u_r) \qquad (17)$$

where u* denotes the optimal control, (b) a monotonically decreasing function of T, for $t_2 \geq t_1$

$$V(x_o, 0, t_2; u_2^*) \leq V(x_o, 0, t_1; u_1^*) \quad , \qquad (18)$$

(c) a quadratic function of x_o,

$$V(x_o, 0, T; u^*) = x_o' P(0, T) x_o \quad . \qquad (19)$$

The question of positive semidefiniteness of the polynomial $C(\omega^2)$ is now reduced to a question regarding the asymptotic behavior of the optimal cost. Therefore, the answer lies in investigating the behavior by some convenient technique -- The Riccati equation [7]. By constancy of the problem, we may investigate the cost $x' P(-T, 0)x$ instead of $x' P(0, T)x$.

$$V(x_o, 0, \infty; u^*) = V(x_o, -\infty, 0; u^*) = \lim_{t \to -\infty} V(x_o, t, 0; u^*)$$
$$V(x_o, t, 0; u^*) = x_o' P(t, 0) x_o \qquad (20)$$

$$-\frac{dP(t,0)}{dt} = P(t,0)F + F'P(t,0) - P(t,0)gg'P(t,0) + Q \qquad (21)$$

with initial condition $P(0,0) = P_o$ to be properly chosen for the boundary condition of the optimization problem.

It turns out that there is a convenient way to simultaneously perform the task of expressing P_o for the boundary condition x(T) and solving the matrix Riccati equation -- by the following so-called "Hamiltonian system of differential equation." This is known in optimization theories [7][3]:

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = A \begin{bmatrix} x \\ p \end{bmatrix} \quad , \qquad A = \begin{bmatrix} F & -gg' \\ -Q & F' \end{bmatrix} \tag{22}$$

The solution to Eq. (22) is of the form

$$\begin{bmatrix} x \\ p \end{bmatrix} = \exp At \begin{bmatrix} x_o \\ p_o \end{bmatrix} = \begin{bmatrix} \Phi_{11}(t) & \Phi_{12}(t) \\ \Phi_{21}(t) & \Phi_{22}(t) \end{bmatrix} \begin{bmatrix} x_o \\ p_o \end{bmatrix} \tag{23}$$

The general solution of the Riccati equation, Eq. (21), is, for some matrices B_1 and B_2 depending on the end point boundary condition of the optimal regulartory problem,

$$P(t) = (\Phi_{21}B_1 + \Phi_{22}B_2)(\Phi_{11}B_1 + \Phi_{12}B_2)^{-1} \tag{24}$$

For our special case of x = 0 at terminal time, $B_1 = 0$, $B_2 = I$, so that

$$P(t,0) = \Phi_{22}(t)\Phi_{12}^{-1}(t) \quad . \tag{25}$$

NOTE (26)

Instead of using Eqs. (23) and (25), the following procedure is used [3]. One first obtains the transition matrix

$$\Phi(-\Delta t) = \begin{bmatrix} \Phi_{11}(-\Delta t) & \Phi_{12}(-\Delta t) \\ \Phi_{21}(-\Delta t) & \Phi_{22}(-\Delta t) \end{bmatrix} = \exp A(-\Delta t)$$

by, for example, a series expansion of exp $A(\Delta t)$ using some small value of Δt. One then performs the following iteration

$$P(-\Delta t, 0) = \Phi_{22}(-\Delta t)\Phi_{12}^{-1}(-\Delta t)$$

$$P(-2\Delta t, 0) = [\Phi_{21}(-\Delta t) + \Phi_{22}(-\Delta t)P(-\Delta t,0)][\Phi_{11}(-\Delta t) + \Phi_{12}(-\Delta t)P(-\Delta t,0)]^{-1}$$

$$P[-(k+1)\Delta t, 0)] = [\Phi_{21}(-\Delta t) + \Phi_{22}(-\Delta t)P(-k\Delta t, 0)][\Phi_{11}(-\Delta t) + \Phi_{12}(-\Delta t)P(-k\Delta t, 0)]^{-}$$

step taking care to symmetrize the P matrix after each to keep the numerical round-off errors from propagating. The process terminates either when condition Eq. (28) occurs or when P has reached a stable value, Eq. (29), by some convergence test, for example as in [3],

$$\sum_{i=1}^{n} |p_{ii}(-k\Delta t - \Delta t) - p_{ii}(-k\Delta t)| / \sum_{i=1}^{n} |p_{ii}(-k\Delta t - \Delta t)| < \text{Some given } \epsilon$$

where $P_{22}(-k\Delta t)$ is the (n)th element of $P(-k\Delta t, 0)$.

This matrix $P(t, 0)$ gives the positive semidefiniteness check and the spectral factorization by the corollary as follows.

Although $f(x_o)$ of Eq. (17) depends on u_r, for definiteness,

Let $u_r(t) = a$, a constant, from t_o (negative constant) to 0. Integrate Eq. (13) with $x(t_o) = 0$, find $x_o = x(0)$, and calculate

$$f(x_o) = -\tfrac{1}{2}\int_{t_o}^{0} (a^2 + x'(t)Qx(t))dt \qquad (27)$$

By Eq. (18), $x_o' P(t, 0)x_o$ decreases monotonically as $t \to -\infty$. We check $x_o' P(t, 0)x_o$ frequently, for example, after each iteration in Eq. (26), for the following conditions.

If $x_o' P(t, 0)x_o < f(x_0)$ for some $t < 0$, then $C(\omega^2)$ is not positive semi-definite and no spectral factorization of $(C(\omega^2)$ is possible. (28)

The Spectral Factorization

Now we allow t to run in the negative direction until the following matrix is obtained.

$$P = \lim_{t \to -\infty} P(t, 0) \qquad . \qquad (29)$$

The spectral factorization is given by

$$C(\omega^2) = [\psi(i\omega) + k(i\omega)]\,[\psi(-i\omega) + k(-i\omega)] \tag{30}$$

where

$$k(s) = k_n s^{n-1} + k_{n-1} s^{n-2} + \ldots + k_1 \tag{31}$$

and

$$[k_1 k_2 \cdots k_n] = k' = g'P = [0 \cdots 0\ 1]\,P \tag{32}$$

To prove Eqs. (30-32), we start from the steady state solution of Eq. (21), $-PF - F'P + Pgg'P = Q$. Adding and substracting $i\omega P$ from the left-hand side, using Eqs. (12) and (32), premultiplying both sides by $g'\Phi'$, and postmultiplying by $\bar{\Phi}g$ using Eq. (32), we have

$$+g'\Phi k + k'\bar{\Phi}g + (k'\Phi g)(k'\bar{\Phi}g) = g'\Phi Q\bar{\Phi}g$$

Adding one to both sides and using Eqs. (6) and (16), we have

$$(1 + k'\Phi g)\,(1 + k'\bar{\Phi}g) = 1 + g'\Phi Q\bar{\Phi}g = \frac{C(\omega^2)}{\psi(i\omega)\psi(-i\omega)}$$

$$\frac{[\psi(i\omega) + k(i\omega)]\,[\psi(-i\omega) + k(-i\omega)]}{\psi(i\omega)\psi(-i\omega)} = \frac{C(\omega^2)}{\psi(i\omega)\psi(-i\omega)}$$

which proves Eq. (30). That it is a spectral factorization follows from the fact that the optimal closed loop system of Theorem 1 cannot be unstable. The characteristic polynomial $\psi(s) + k(s)$ has no roots with positive real parts due to the boundary condition of theorem 1.

Example

The following simple example is intended only to show the reader that the method indeed works. The power of the method is more fully used when the degree of $C(\omega^2)$ becomes high.

We wish to check the polynomial $C(\omega^2) = \omega^2 + 1$. The following stable dynamical system is arbitrary chosen $F = -2$, $g = 1$, which is a controllable system because the canonical form of Eq. (5a) is used. This system has the characteristic polynomial $\psi(s) = s + 2$. Therefore, $\psi(i\omega)\psi(-i\omega) = (i\omega+2)(-i\omega+2) = \omega^2 + 4$ which gives, by Eq. (6) and Eq. (8), $Q = -3$. Using Eq. (23), we have

$$A = \begin{bmatrix} -2 & -1 \\ 3 & +2 \end{bmatrix} \qquad e^{At} = \frac{1}{2}\left[\begin{array}{c:c} -e^{t} + 3e^{-t} & -e^{t} + e^{-t} \\ \hdashline 3e^{t} - 3e^{-t} & 3e^{t} - e^{-t} \end{array}\right]$$

By Eqs. (25), (32) and (30), we have

$$P = \lim_{t \to -\infty} \frac{3e^{t} - e^{-t}}{-e^{t} + e^{-t}} = \lim_{t \to -\infty} \frac{3e^{2t} - 1}{-e^{2t} + 1} = -1. \qquad k = Pg = -1.$$

$$\begin{aligned} C(\omega^2) &= [\psi(i\omega) + k(i\omega)][\psi(-i\omega) + k(-i\omega)] \\ &= (i\omega + 2 - 1)(-i + 2 - 1) = (i\omega + 1)(-i\omega + 1). \end{aligned}$$

CONCLUSIONS

We have shown that the positive semidefiniteness of $C(\omega^2)$ may be checked by sequentially applying Eqs. (3), (4), (5), (5a), (6), (8), (27), (22), and (26), the spectral factorization is obtained by applying Eqs. (32, 31, 30). The general idea is to "imbed" the polynomial into an optimization problem and to watch the solution of the Riccati equation -- if the solution becomes "smaller" than a certain pre-calculated number -- then $C(\omega^2)$ is not positive semidefinite; if the solution stabilizes above this pre-calculated number, then the solution gives a spectral factorization. This process is facilitated by the fact that the convergence is monotonic. Also, the rate of convergence may be made arbitrary fast by choosing the system into which $C(\omega^2)$ is imbedded to have fast dynamics, at the expense of accuracy due to the subtraction of two large numbers in Eq. (30).

REFERENCES

[1] Kalman, R. E., "When is a Linear Control System Optimal?", Journal of Basic Engineering, Trans ASME, March 1964, pp 51-60.

[2] Popov, V. M., "Hyperstability and Optimality of Automatic Systems with Several Control Functions," Revue Roumaine Des Sciences Techniques Electrotechnique, Et Energetique, T. 9,4, Bucarest, 1964, pp 629-690.

[3] Kalman, R. E. and Englar, T. S., "A User's Manual for ASP C," RIAS report for NASA, Baltimore, Maryland, 1965.

[4] Guillemin, E. A., "Synthesis of Passive Networks," John Wiley and Sons, New York, 1957.

[5] Kalman, R. E., "Lyapunov Functions for the Problem of Lur'e in Automatic Control," Proc U. S. National Academic of Science, Vol 49, 1963, pp 201-205.

[6] Newton, G. C., Gould, L. A., and Kaiser, J. F., "Analytical Design of Linear Feedback Systems," John Wiley and Sons, Inc., New York, 1957.

[7] Kalman, R. E., "The Theory of Optimal Control and the Calculus of Variations," Proc RAND-Univ California Symp on Optimization Theory, 1960; U of California Press, 1963, Chapt 16.

[8] Chang, S. S. L., "Synthesis of Optimum Control Systems," McGraw-Hill Book Co., Inc., New York, 1961.

DEVELOPMENT AND APPLICATIONS OF A SYSTEM THEORY CRITERION FOR RATIONAL POSITIVE REAL MATRICES

B. D. O. ANDERSON
Stanford University
Stanford, California

ABSTRACT

An algebraic description of the constraints imposed on a matrix of rational functions by positive reality is presented. The application of this result to the following problem is then discussed:

(a) Spectral factorization by algebraic means;
(b) Characterization of the optimality of linear systems in terms of a closed loop transfer function matrix;
(c) Stability of control systems with multiple memoryless nonlinearities (extended Popov criterion)
(d) Passive network synthesis of positive real immittance matrices by algebraic means.

INTRODUCTION

In 1963, Kalman [1] suggested an alternative viewpoint for the characterization of a restricted class of rational positive real matrices. This characterization was markedly different from those generally appearing in the literature, particularly the literature of network theory, e.g. [2], in that the characterization was in algebraic rather than analytic terms. Similar ideas appeared in a little-known paper by Popov [3]. A neat application of the ideas of [1] appeared in [4], dealing with the Lur'e problem of automatic control and this indirectly suggested the possible significance and power of the algebraic description of the positive real property.

The aim of this paper is to review an extension of Kalman's original result [1] to encompass all positive real matrices, and to summarize a number of problems, the solutions to which have been obtained by the extension. Results in general will be stated, rather than proved, proofs being available in the references.

Positive real matrices arise naturally in network theory, where one of the key results is that a network composed of an arbitrary interconnection of passive resistors, inductors, capacitors, transformers and gyrators has a port impedance (or admittance) which, assuming it exists, is rational and positive real. To show however that positive real matrices are important for other than network theoretic reasons is an ancillary aim of this paper.

SYSTEM THEORY DESCRIPTION OF THE POSITIVE REAL PROPERTY

We term a matrix $Z(s)$ of functions of the complex variable s positive real if it satisfies the following three conditions:

(a) The elements of Z are analytic for $\operatorname{Re} s > 0$

(b) $Z^*(s) = Z(s^*)$ for $\operatorname{Re} s > 0$

(c) $Z'(s^*) + Z(s)$ is nonnegative definite for $\operatorname{Re} s > 0$

Without further comment it will henceforth be assumed that the elements of Z are rational in s. With no loss of generality, we may then restrict consideration to those $Z(s)$ with no pole at infinity, for otherwise, see [2], it is possible to write $Z(s) = sL + Z_1(s)$ where $Z_1(\infty)$ is finite and $Z_1(s)$ is positive real. We shall suppose often that $Z(s)$ has a minimal realization $\{F, G, H, J\}$, [5]. In other words $Z(s)$ is related to the four constant matrices F, G, H and J by the equation

$$Z(s) = J + H'(sI-F)^{-1}G \tag{1}$$

and F, being square, is of minimal dimension.

The algebraic characterization of the positive real property, a proof of which may be found in [6] or [7], is as follows:

Theorem 1. Let $Z(s)$ be a matrix of rational transfer functions such that $Z(\infty)$ is finite, and the elements of Z have poles which lie in $\operatorname{Re} s < 0$ or which are simple on $\operatorname{Re} s = 0$. Let $\{F, G, H, J\}$ be a minimal realization for Z. Then $Z(s)$ is positive real if and only if there exists a symmetric positive definite P, and matrices L and $W_o(\infty)$ such that

$$PF + F'P = -LL' \tag{2}$$

$$PG = H - L\,W_o(\infty) \tag{3}$$

$$W_o'(\infty)W_o(\infty) = J + J' \tag{4}$$

It is of interest to appreciate how the matrices P, L and $W_o(\infty)$ arise. While [6] or [7] contain a full discussion, we point out here that these matrices derive from carrying out a spectral factorization, as discussed by e.g. Youla [8]. As is shown in this reference, given a positive real $Z(s)$ there exists a matrix $W_o(s)$, uniquely defined within left multiplication by an arbitrary constant orthagonal matrix, such that

(a) $Z(s) + Z'(-s) = W_o'(-s)\,W_o(s)$ (5)

(b) The elements of $W_o(s)$ are analytic for $\operatorname{Re} s \geq 0$

(c) $W_o(s)$ has constant rank in $\operatorname{Re} s \geq 0$

At least in the case when $Z(s)$ has no $j\omega$-axis poles, it is true [6, 7] that the L and $W_o(\infty)$ of (2), (3) and (4) define, together with F and G, a minimal realization $\{F, G, L, W_o(\infty)\}$ of this $W_o(s)$ by

$$W_o(s) = W_o(\infty) + L'(sI-F)^{-1}G \tag{6}$$

Equation 2 can then be looked on as defining the matrix P, while equation 3 turns out to be automatically satisfied because of (2), (4), (5) and (6). (That $Z(s)$ and $W(s)$ have the same poles is not hard to establish, from which it can then be shown that W_o and Z may be taken with the same resolvent matrix. To show that G in the realization $\{F, G, H, J\}$ can be incorporated in a minimal realization for $W_o(s)$ is highly nontrivial, except in the case where Z and W_o are scalar functions.)

The determination of $W_o(\infty)$, L and P is possible in at least three ways. First, one could attempt directly to solve (2), (3) and (4), presumably with some iterative procedure. Second, one could use the constructive techniques of Youla [8] or Davis [9] to determine $W_o(s)$ satisfying (5); using this $W_o(s)$, L and $W_o(\infty)$ would then be found from (6), and finally P from the linear equation (2). Third, one can consider:

SPECTRAL FACTORIZATION BY ALGEBRAIC MEANS

The spectral factorization problem is as follows:

Given $\Phi(s)$ satisfying

$$\Phi'(-s) = \Phi(s) \tag{7}$$

$$\Phi(j\omega) \geq 0 \tag{8}$$

find a matrix $W(s)$ with elements analytic in $\text{Re } s \geq 0$ such that

$$\Phi(s) = W'(-s)\, W(s) \tag{9}$$

Often, one must find that particular matrix $W(s)$, call it $W_o(s)$, which has constant rank in the right half plane.

It is shown in [10] that the general factorization problem can be reduced via simple matrix transformations to that of factoring $\Phi(s)$ such that $\Phi(\infty)$ is nonsingular and such that $\Phi(s)$ can be written as

$$\Phi(s) = Z(s) + Z'(-s) \tag{10}$$

where $Z(s)$ is positive real with elements whose poles are all in $\text{Re } s < 0$.

The following theorem is then demonstrated in [10]:

Theorem 2. Let $Z(s)$ in (10) have a minimal realization $\{F, G, H, J\}$. Then matrices $W(s)$ satisfying (9) are defined by realizations

$\{F, G, (PG-H)(J+J')^{-1/2}, (J+J')^{1/2}\}$ where P satisfies the quadratic matrix equation

$$PF + F'P = -(PG-H)(J+J')^{-1} (PG-H)' \tag{11}$$

Moreover $W_o(s)$ is defined by that $P = P_o$ which is also such that the eigenvalues of $F + G(J+J')(PG-H)'$ have negative real part. Finally, P_o is unique, is positive definite symmetric, and is the matrix P of equations (2) and (3).

The solution of (11) is also discussed in [10], where the following is derived.

Theorem 3. Consider the matrix

$$M = \begin{bmatrix} F' + H(J+J')^{-1}G' & H(J+J')^{-1}H' \\ G(J+J')^{-1}G' & -F + G(J+J')^{-1}H' \end{bmatrix} \tag{12}$$

of dimension $2n \times 2n$ if F is of dimension n. Let $c_1, c_2, \ldots c_{2n}$ be eigenvectors (generalized if necessary) of M, in arbitrary order, and let

$$\begin{bmatrix} A \\ \cdots \\ B \end{bmatrix} = \left[c_1 c_2 \ldots c_n\right] \tag{13}$$

define $n \times n$ matrices A and B. Then if B^{-1} exists, a solution of (11) is AB^{-1}. Moreover, M has precisely n eigenvalues of positive real part, and if $c_1, \ldots c_n$ correspond to these, P_o is given by AB^{-1}.

The extended factorization procedures of Youla or Davis can thus be replaced by the none-too-easy procedure of finding eigenvectors of a matrix of dimension in essence determined by the complexity of $\Phi(s)$. It also seems likely that solutions to (11) can be determined as limiting values of the solutions to an associated Riccati equation.

OPTIMALITY OF LINEAR SYSTEMS

A well-known [11] and useful procedure of optimal control is to optimize the system, assumed completely controllable,

$$\dot{x} = Fx + Gu \tag{14}$$

with performance index

$$V(x_o) = \int_{t_o}^{\infty} (x'Qx + u'u)\, dt \tag{15}$$

(with $[F, Q^{1/2}]$ assumed completely observable). This optimization leads to the determination of a control law

$$u = -Kx \tag{16}$$

where K is a constant matrix. The following frequency domain characterization of optimality may then be given [12].

Theorem 4. Suppose the system (14) is optimized with respect to the performance index (15) with control (16). Then

$$[I + G'(-j\omega I-F')^{-1}K]\,[I + K'(j\omega I-F)^{-1}G] - I \geq 0 \tag{17}$$

Conversely if for some linear control law K for the system (14), the inequality

$$[I + G'(-j\omega I-F')^{-1}K]\,[I + K'(j\omega I-F)^{-1}G] - I > 0 \tag{18}$$

holds, then there exists a Q and corresponding performance index (15) such that (16) is the associated optimal control law.

This theorem extends an earlier result [13] applicable to single-input systems to the multiple input case. The first part of the theorem follows as a simple extension of the ideas of [13], while the second half requires for its proof the application of the algebraic positive real description. For details, see [12]. Extensions of this result to a corresponding stochastic control problem appears in [14].

STABILITY OF CONTROL SYSTEMS WITH MULTIPLE MEMORYLESS NONLINEARITIES

The material discussed here summarizes results in [15]. We consider a control system consisting of a linear plant with stable transfer function matrix W(s), and with the input to the plant $-\psi$, an n-vector, and output of the plant σ, also an n-vector. The input is derived by feeding back the output, i.e. $\psi = \psi(\sigma)$, with the restrictions

$$0 \leq \frac{\psi_i}{\sigma_i} \leq k_i > 0 \qquad (i=1,2, \ldots n) \tag{19}$$

Thus we are generalizing the usual Popov problem [4, 16] to n nonlinearities. The principal result is as follows:

Theorem 5. With $K = \text{diag}\,\{k_i\}$, if there exist nonnegative constants α and β, at least one positive, such that $K^{-1} + (\alpha+\beta s)\,W(s)$ is positive real, then the system is stable for arbitrary nonlinearities satisfying (19).

We remark that the proof of this theorem in [15] rests on applying the algebraic positive real characterization to $K^{-1} + (\alpha + \beta s)\,W(s)$, from which a 'P' matrix (see equations (2), and (3)) is derived. A lyapunov function may then be formed which includes the term $x'\,P\,x$, and using this lyapunov function stability can be demonstrated. As has often proved the case in past stability problems, the difficult problem is to define the lyapunov function, the easy problem is to verify that it satisfies the requisite conditions.

PASSIVE NETWORK SYNTHESIS

In this section, the classical network synthesis problem is tackled: given a positive real $Z(s)$, find a network of resistors, inductors, capacitors, transformers and gyrators with $Z(s)$ as its impedance matrix. We give a new solution to this problem in the following theorem, see [17] for proof:

Theorem 6. Let $Z(s)$ be an $m \times m$ positive real matrix, and let $\{F, G, H, J\}$ be a minimal realization. Let P be the matrix of Theorem 1. Suppose F has a dimension n. Then the impedance matrix

$$Z_c = \begin{bmatrix} J & -H'P^{-1/2} \\ P^{1/2}G & -P^{1/2}F\,P^{-1/2} \end{bmatrix} \tag{20}$$

is nonnegative definite, and thus may be synthesized by an $(n+m)$-port network of passive resistors, gyrators and transformers. Moreover, if the last n ports of this network are terminated in unit inductors, the impedance offered at the remaining m ports is precisely $Z(s)$. Finally, the resulting synthesis of $Z(s)$ uses the minimal number of resistors and inductors.

Note that $P^{1/2}$ is well defined because P is positive definite symmetric. Note also that the synthesis, though apparently algebraic, does rely in some way on the spectral factorization idea to obtain P, and the spectral factorization procedure itself does require analytical work to carry it through.

Extensions of the spectral factorization ideas and the ideas behind Theorem 6 lead to the investigation of equivalent networks, i.e., the characterization of all, rather than one, networks realizing a prescribed $Z(s)$. This is currently under investigation. Recently, [18], others have started to obtain results along these lines; a particularly pertinent problem is the determination of networks containing no gyrators which synthesize a symmetric positive real $Z(s)$.

REFERENCES

1. R. E. Kalman, "On a New Characterization of Linear Passive Systems," Proceedings of the First Allerton Conference on Circuit and System Theory, University of Illinois, November 1963, pp. 456-470.
2. R. W. Newcomb, Linear Multiport Synthesis, McGraw Hill, New York, 1966.
3. V. M. Popov, "Hyperstability and Optimality of Automatic Systems with Several Control Functions", Rev. Roum. Sci. Tech., Electrotechn. et Energ., Vol. 9, No. 4, 1964, pp. 629-690.
4. R. E. Kalman, "Lyapunov Functions for the Problem of Lur'e in Automatic Control," Proceedings of the National Academy of Sciences, Vol. 49, No. 2, February 1963, pp. 201-205.
5. R. E. Kalman, "Mathematical Description of Linear Dynamical Systems," Journal of the Society for Industrial and Applied Mathematics Control, Series A, Vol. 1, No. 2, 1963, pp. 152-192.
6. B. D. O. Anderson, "A Control Theory Viewpoint of Positive Reality," Rept. No. SEL-66-009 (TR No. 6558-4), Stanford Electronics Laboratories, Stanford, California, February 1966.
7. B. D. O. Anderson, "A System Theory Criterion for Positive Real Matrices," SIAM Journal of Control, to appear.
8. D. C. Youla, "On the Factorization of Rational Matrices," IRE Transactions on Information Theory, Vol. IT-7, No. 3, July 1961, pp. 172-189.
9. M. C. Davis, "Factoring the Spectral Matrix," IEEE Transactions on Automatic Control, Vol. AC-8, No. 4, October 1963, pp. 296-305.
10. B. D. O. Anderson, "Spectral Factorization by Algebra", Rept. TR 6560-6, Stanford Electronics Laboratories, Stanford, Calif., September 1966.
11. M. Athans and P. L. Falb, Optimal Control, McGraw Hill, New York, 1966.

12. B. D. O. Anderson, "The Inverse Problem of Optimal Control," Rept. No. SEL-66-038 (TR No. 6560-3), Stanford Electronics Laboratories, Stanford, California, May 1966.

13. R. E. Kalman, "When is a Linear Control System Optimal?", Journal of Basic Engineering, Transactions of the American Society of Mechanical Engineers, Series D, Vol. 86, March 1964, pp. 1-10.

14. B. D. O. Anderson, "The Testing of Optimality for Linear Systems", International Journal of Control, to appear.

15. B. D. O. Anderson, "Stability of Control Systems with Multiple Nonlinearities," Journal of the Franklin Institute, Vol. 281, No. 9, September 1966, to appear.

16. V. N. Popov, "Absolute Stability of Nonlinear Systems of Automatic Control," Avtomatika i Telemekhnika, Vol. 22, August 1961, pp. 961-979 and Automation and Remote Control, Vol. 22, No. 8, March 1962, pp. 857-875.

17. B. D. O. Anderson and R. W. Newcomb, "Impedance Synthesis via State Space Techniques," Rept. No. SEL-66-024, (TR No. 6558-5), Stanford Electronics Laboratories, Stanford, California, April 1966.

18. E. S. Kuh, D. M. Layton and J. Tow, "Network Analysis and Synthesis via State Variables", Memorandum No. ERL-M169, Electronics Research Laboratory, University of California, Berkeley, July 1966.

NUMBER OF TREES IN A CASCADE OF TWO-PORT NETWORKS

B. R. MYERS
University of Notre Dame
Notre Dame, Indiana

INTRODUCTION

Recurrence formulas for the number of trees and of port terminal-pair 2-trees in a cascade of identical two-port networks have been presented by W. Gruszczynski.[1] The purpose of the present paper is to present the generalized recurrence formulas which pertain when the networks in the cascade are non-identical; not only for the trees and port terminal-pair 2-trees, but also for the terminal 3-trees. These formulas are derived from counting series which provide an insight, not apparent from the recurrence formulas themselves, into the composition of the various tree structures. The generalized recurrence formulas, and the series from which they derive, reduce readily to the simpler forms which pertain in the special case when the networks in the cascade are identical.

DEFINITIONS

Consider a cascade C_n of an arbitrary two-port networks $G_1, G_2, \ldots, G_n$, connected in that order, as shown in Figure 1; with $(0,0')$ the input terminal pair of G_1, and (k,k'), $k=1,2,\ldots,n$, the output terminal pair of network G_k in C_n. Suppose g_k is the graph of G_k, and let

$$\begin{aligned} n_k &= \text{an arbitrary tree of } g_k, \\ x_k &= \text{an arbitrary 2-tree } T_{(k-1),(k-1)'} \text{ of } g_k, \\ y_k &= \text{an arbitrary 2-tree } T_{k,k'} \text{ of } g_k, \\ z_k &= \text{an arbitrary 3-tree } T_{(k-1),k,(k-1)'k'}, \\ &\quad T_{(k-1),k',(k-1)'k}, T_{(k-1)k,(k-1)',k'}, \\ &\quad \text{or } T_{(k-1)k', (k-1)',k} \text{ of } g_k; \end{aligned} \tag{1}$$

with

$$\begin{aligned} N_k &= \text{the number of trees } n_k \text{ in } g_k, \\ X_k &= \text{the number of 2-trees } x_k \text{ in } g_k, \\ Y_k &= \text{the number of 2-trees } y_k \text{ in } g_k, \\ Z_k &= \text{the number of 3-trees } z_k \text{ in } g_k. \end{aligned} \tag{2}$$

For any such cascade C_{k_j} comprising the k_j two-ports $G_{k_1}, G_{k_2}, \ldots, G_{k_j}$, connected in that order, with $((k_1-1),(k_1-1)')$ the input terminal pair of G_{k_1}, and (k_g, k_g'), $g=1,2,\ldots,j$, the output terminal pair of G_{k_g}, let

$$
\begin{aligned}
N_{k_1,k_2,\ldots,k_j} &= \text{the number of trees in } C_{k_j}, \\
X_{k_1,k_2,\ldots,k_j} &= \text{the number of 2-trees } T_{(k_1-1),(k_1-1)'} \text{ in } C_{k_j}, \\
Y_{k_1,k_2,\ldots,k_j} &= \text{the number of 2-trees } T_{k_j,k_j'} \text{ in } C_{k_j}, \\
Z_{k_1,k_2,\ldots,k_j} &= \text{the number of 3-trees } T_{(k_1-1),k_j,(k_1-1)'k_j'}, \\
&\quad T_{(k_1-1),k_j',(k_1-1)'k_j}, T_{(k_1-1)k_j,(k_1-1)',k_j'}, \\
&\quad \text{or } T_{(k_1-1)k_j', (k_1-1)',k_j}.
\end{aligned}
\tag{3}
$$

For use when the k_j networks in C_{k_j} are identical, the following notation for the quantities in (3) are introduced:

$$
\begin{aligned}
N_{k_1,k_2,\ldots,k_j}\Big|_{G_{k_1} = G_{k_2} = \ldots = G_{k_j}} &\triangleq N_{((j))} \\
X_{k_1,k_2,\ldots,k_j}\Big|_{G_{k_1} = G_{k_2} = \ldots = G_{k_j}} &\triangleq X_{((j))} \\
Y_{k_1,k_2,\ldots,k_j}\Big|_{G_{k_1} = G_{k_2} = \ldots = G_{k_j}} &\triangleq Y_{((j))} \\
Z_{k_1,k_2,\ldots,k_j}\Big|_{G_{k_1} = G_{k_2} = \ldots = G_{k_j}} &\triangleq Z_{((j))}
\end{aligned}
\tag{4}
$$

THE TREES

A counting series for the trees of the n-cascade C_n of Figure 1 then evolves by considering any tree T_n of C_n. T_n comprises an n-cascade of subgraphs t_k, $k=1,2,\ldots,n$, each either a tree n_k, a 2-tree x_k or y_k, or a 3-tree z_k, of the respective network graphs g_k in C_n, as defined by (1). Denote by H_j, $0 \leqq j \leqq (n-1)$, the sub-cascade of T_n comprising all those t_k for which $k \leqq j$; with $\overline{H}_j = (T_n - H_j)$ the sub-cascade which complements H_j in T_n. Since T_n is a tree, there is a unique path between the pair of terminals (j,j') which are common to H_j and $\overline{H}_j$. Hence, either one or the other of $H_j, \overline{H}_j$, but not both, is a tree. In particular, $\overline{H}_o$ is always a tree. And by end-to-end enumeration symmetry of the n-cascade, the set of all possible configurations H_j for each value of j in the range $1 \leqq j \leqq (n-1)$ contains at least one tree T_j. Further, for $j=(n-1)$, the subgraph t_n in the n-cascade is necessarily a 2-tree of the type x_k, $k=n$, in (1).

The number $N_{1,2,\ldots,n}$ of all possible trees T_n in C_n is therefore given by

$$N_{1,2,\ldots,n} = \sum_{j=0}^{(n-1)} F_j \tag{5}$$

where (a) F_o is the number of configurations in which $\overline{H}_o$ is a tree and H_j, $1 \leqq j \leqq (n-1)$, is not; (b) each F_j for which $1 \leqq j \leqq (n-2)$, is the number of configurations in which H_j is a tree and H_{j+k}, $1 \leqq k \leqq n-(j+1)$, is not; and (c) $F_{(n-1)}$ is the number of configurations in which $H_{(n-1)}$ is a tree.

Using the quantities defined in (2), it is then not difficult to see that the respective topologies which necessarily hold dictate that:

$$F_o = \prod_{q=1}^{(n-1)} Y_q N_n \tag{6}$$

$$F_j = N_{1,2,\ldots,j} Z_{(j+1)} \prod_{p=(j+2)}^{(n-1)} Y_p N_n, \text{ for } 1 \leqq j \leqq (n-3) \tag{7}$$

$$F_{(n-2)} = N_{1,2,\ldots,(n-2)} Z_{(n-1)} N_n \tag{8}$$

$$F_{(n-1)} = N_{1,2,\ldots,(n-1)} N_n \tag{9}$$

For example, (7) follows from the fact that H_j is a tree comprising the subgraphs $t_1, t_2, \ldots, t_j$; subgraph $t_{(j+1)}$ is necessarily a 3-tree of the type z_k, $k=(j+1)$; each of the subgraphs $t_{(j+2)}, t_{(j+3)}, \ldots, t_{(n-1)}$ is necessarily a 2-tree of the type y_k; and subgraph t_n is necessarily a tree of the type n_k.

Finally, therefore, from (5) through (9), the counting series for the number of trees in a cascade of arbitrary two-port networks, Figure 1, is:

$$N_o = 0 \; ; \; N_1 \equiv N_1 \; ;$$

and for $n \geqq 2$,

$$\begin{aligned} N_{1,2,\ldots,n} = {} & N_n \prod_{q=1}^{(n-1)} Y_q \\ & + N_n \sum_{j=1}^{(n-3)} N_{1,2,\ldots,j} Z_{(j+1)} \prod_{p=(j+2)}^{(n-1)} Y_p \\ & + N_n Z_{(n-1)} N_{1,2,\ldots,(n-2)} \\ & + X_n N_{1,2,\ldots,(n-1)} \end{aligned} \tag{10}$$

A recurrence formula follows from (10), with very little algebraic manipulation, thus

$$N_o = 0 \; ; \; N_1 \equiv N_1 ;$$

and for $n \geqq 2$,

$$\begin{aligned} N_{1,2,\ldots,n} N_{(n-1)} &= \{N_{(n-1)} X_n + Y_{(n-1)} N_n\} N_{1,2,\ldots,(n-1)} \\ &\quad + \{N_{(n-1)} Z_{(n-1)} - X_{(n-1)} Y_{(n-1)}\} N_{1,2,\ldots,(n-2)} N_n \end{aligned} \tag{11}$$

If the two-port networks G_k, $k=1,2,\ldots,n$, in the n-cascade C_n of Figure 1 are identical, then (10) reduces to (12), below, and (11) reduces to (13).

$$N_{((n))} = X_1 N_{((n-1))} + N_1 Z_1 \sum_{j=1}^{(n-2)} N_{((j))} Y_1^{n-(j+2)} + N_1 Y_1^{(n-1)}$$

for $n \geqq 1$; where (12)

$$N_{((1))} \equiv N_1 \; ; \; N_{((0))} = 0.$$

$$N_{((n))} = (X_1 + Y_1) N_{((n-1))} + (N_1 Z_1 - X_1 Y_1) N_{((n-2))}$$

for $n \geqq 2$; where (13)

$$N_{((1))} \equiv N_1 ; \; N_{((0))} = 0$$

Equation (13) is recognized as Gruszczynski's result[1].

Of course, though not noted by Gruszczynski, equations (12) and (13) apply not only to cascades of topologically identical networks but also to cascades made up, in whole or part, of non-identical networks G_k, $k=1,2,..,n$, each of which has the same set of values N_k, X_k, Y_k, Z_k; for example, to a cascade made up of either or both of the two networks shown in Figures 2(a) and 2(b), in any order. The reader may confirm that $N_k=3$, $X_k=8=Y_k$, and $Z_k=21$, for each of these networks.

THE PORT TERMINAL-PAIR 2-TREES

The several counting series and recurrent relations, below, for the number $X_{1,2,\ldots,n}$, and, when the networks are identical, $X_{((n))}$, of the input port terminal-pair (0,0') 2-trees for the n-cascade of two-port networks C_n of Figure 1, are given without proof; since their derivation follows along exactly those lines which lead to equations (10) through (19) for the trees of C_n. Formulas for the output port terminal-pair (n,n') 2-trees of C_n, in number $Y_{1,2,\ldots,n}$ or $Y_{((n))}$, are not given; since, of course, they have exactly the same form as those for $X_{1,2,\ldots,n}$ and $X_{((n))}$.

First, a counting series for $X_{1,2,\ldots,n}$ in the general case when the networks in C_n are all different:

$$
\begin{aligned}
X_{1,2,\ldots,n} = {} & N_n Z_1 \prod_{q=2}^{(n-1)} Y_q \\
& + N_n \sum_{j=1}^{(n=3)} X_{1,2,\ldots,j} Z_{(j+1)} \prod_{p=(j+2)}^{(n-1)} Y_p \\
& + N_n Z_{(n-1)} X_{1,2,\ldots,(n-2)} \\
& + X_n X_{1,2,\ldots,(n-1)}
\end{aligned}
\tag{14}
$$

for $n \geq 1$; where

$$X_1 \equiv X_1 \;;\; X_o = 1 \;;\; Z_o = 0.$$

The reader may note the similarity in form of (14) and (10); correspondingly with (15), below, the recurrence relation derived from (14), and (11), the recurrence relation for trees which was derived from (10). (The first and second terms in (14) could, incidentally, be combined by putting j=0 as the lower limit on j in the summand in the second term.)

$$
\begin{aligned}
X_{1,2,\ldots,n} = {} & \{X_n + Y_{(n-1)}\} X_{1,2,\ldots,(n-1)} + \{Z_{(n-1)} N_n \\
& - X_{(n-1)} Y_{(n-1)}\} X_{1,2,\ldots,(n-2)}
\end{aligned}
$$

for $n \geq 2$; where (15)

$$X_1 \equiv X_1 \;;\; X_o = 1.$$

Another counting series for the 2-trees is readily derived from (11) and (14) as

$$
\begin{aligned}
X_{1,2,\ldots,n} = {} & \prod_{q=1}^{n} X_q \\
& + \sum_{j=0}^{(n-2)} N_{(j+2),(j+3),\ldots,n} Z_{(j+1)} \prod_{p=0}^{j} X_p
\end{aligned}
\tag{16}
$$

for $n \geq 2$; where

$$X_1 \equiv X_1 \;;\; X_o = 1.$$

An interesting relation follows from (11) and (14), as

$$(N_1 X_{1,2,\ldots,n} - X_1 N_{1,2,\ldots,n}) = (N_1 Z_1 - X_1 Y_1) N_{2,3,\ldots,n} \tag{17}$$

for $n \geq 2$.

In the special case when all networks in the cascade are either identical or else have the same set of values for N_k, X_k, Y_k, Z_k, equations (14), (15), (16), and (17) reduce respectively to the following:

$$X_{((n))} = X_1 X_{((n-1))} + N_1 Z_1 \sum_{j=0}^{(n-2)} X_{((j))} Y_1^{n-(j+2)}$$

for $n \geqq 2$; where (18)

$$X_{((1))} \equiv X_1; \; X_{((0))} = 1 .$$

$$X_{((n))} = (X_1 + Y_1) X_{((n-1))} + (Z_1 N_1 - X_1 Y_1) X_{((n-2))}$$

for $n \geqq 2$; where (19)

$$X_{((1))} \equiv X_1 ; \; X_{((0))} = 1.$$

$$X_{((n))} = X_1^n + Z_1 \sum_{j=0}^{(n-2)} N_{((n-j-1))} X_1^j$$

for $n \geqq 0$; where (20)

$$X_{((1))} \equiv X_1 ; \; X_{((0))} = 1; \; N_{((1))} \equiv N_1 .$$

$$N_1 X_{((n))} - X_1 N_{((n))} = \{N_1 Z_1 - X_1 Y_1\} N_{((n-1))}$$

for $n \geqq 1$; where $N_{((0))} = 0$. (21)

Equations (19) is recognized as that given by Gruszczynski.[1]

THE TERMINAL 3-TREES

Counting series for the terminal 3-trees are again given without proof, since the technique in their derivation is similar to that for the trees and 2-trees. Whence:

$$Z_{1,2,\ldots,n} = \sum_{j=1}^{n} X_{1,2,\ldots,(n-j)} Z_{(n-j+1)} \prod_{p=(n-j+2)}^{(n+1)} Y_p$$

for $n \geqq 1$; where (22)

$$X_o = 1; \; X_1 \equiv X_1; \; Y_{(n+1)} = 1; \; Z_o = 0; \; Z_1 \equiv Z_1.$$

A recurrence formula follows with very little algebraic manipulation from (22), thus:

$$Z_{(n-1)} Z_{1,2,\ldots,n} = \left\{X_{(n-1)} Z_n + Z_{(n-1)} Y_n\right\} Z_{1,2,\ldots,(n-1)}$$
$$+ \left\{N_{(n-1)} Z_{(n-1)} - X_{(n-1)} Y_{(n-1)}\right\} Z_{1,2,\ldots,(n-2)} Z_n$$

for $n \geqq 2$; where

$$X_1 \equiv X_1 \ ; \ Y_1 \equiv Y_1 \ ; \ Z_o = 0 \ ; \ Z_1 \equiv Z_1 \ ; \ N_1 \equiv N_1 \ . \tag{23}$$

Note that the recurrence formula (23) for the terminal 3-trees has the same general form as (11) and (15); which are the respective corresponding formulas for the trees and input port 2-trees.

When the networks in the cascade are all identical, or when each network has the same set of values for N_k, X_k, Y_k, Z_k, then (22) and (23) reduce respectively to the simpler forms

$$Z_{((n))} = Z_1 \sum_{j=1}^{n} X_{((n-j))} Y_1^{(j-1)}$$

for $n \geqq 1$; where (24)

$$X_{((0))} = 1 \ ; \ X_{((1))} \equiv X_1 \ ; \ Z_{((0))} = 0 \ ; \ Z_{((1))} \equiv Z_1 .$$

$$Z_{((n))} = (X_1+Y_1)Z_{((n-1))} + (N_1Z_1-X_1Y_1)Z_{((n-2))}$$

for $n \geqq 2$; where

$$Z_{((0))} = 0 \ ; \ Z_{((1))} \equiv Z_1 \ . \tag{25}$$

It is interesting to observe from (24) that Z_1 is a factor in the formula for $Z_{((n))}$. Hence the normalized quantity $Z_{((n))}/Z_1$ may be used in counting the terminal 3-trees. The corresponding normalizations $N_{((n))}/N_1$, $X_{((n))}/X_1$, and $Y_{((n))}/Y_1$, respectively for the trees, input port 2-trees, and output port 2-trees, are not, however, strictly valid from the mathematical standpoint; as the reader may check by examination of the appropriate formulas presented previously for $N_{((n))}$, $X_{((n))}$, $Y_{((n))}$.

AN EXAMPLE

Consider the non-symmetric two-port network C_n of Figure 3, with (I,I'), (II,II') the input and output port terminal pairs, respectively.

The numbers N of trees, X of input port 2-trees, Y of output port 2-trees, and Z of terminal 3-trees in C_n are

$$N = 377 \ ; \ X = 233 \ ; \ Y = 233 \ ; \ Z = 144 \tag{26}$$

Thus, even for this relatively simple 8-node, 13-branch network, direct longhand computation of these numbers of trees would be a formidable task. The reader may confirm that the task becomes routine, almost trivial, however, on decomposing the network into a cascade of elementary subnetworks and then applying the formulas given earlier.

REFERENCE

1. W. Gruszczynski, Number of trees in a chain of identical graphs, IEEE Trans. on Circuit Theory, CT-13, June 1966, pp. 221-2.

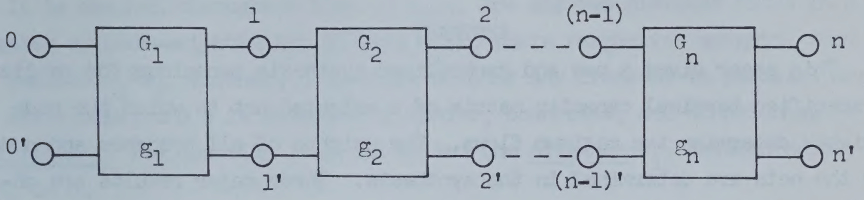

0
0'
G_1
g_1
1
1'
G_2
g_2
2
2'
(n-1)
(n-1)'
G_n
g_n
n
n'

Figure 1. n-Cascade C_n

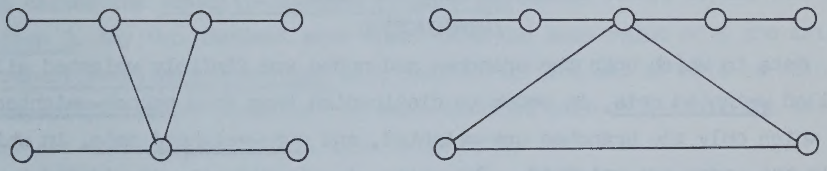


Figure 2

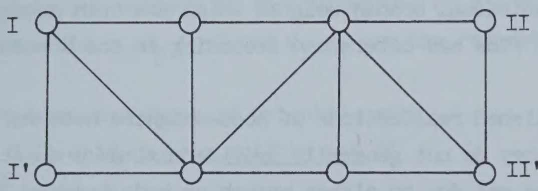


Figure 3. C_n

MINIMAL REALIZATION OF A CLASS OF MAX-FLOW MIN-OUTAGE WEIGHTED NETS

B. R. MYERS
University of Notre Dame
Notre Dame, Indiana

M. A. TAPIA
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

This paper gives a new and generalized synthesis technique for realizing a specified terminal capacity matrix of a weighted net in which the node weights determine the maximum flows. The weights of all branches and nodes in the nets are determined in the synthesis. Three major results are obtained: (1) a useful property, the Capacity-Distance Product Sum, is defined for a fundamental class of weighted nets, (2) the set of node and branch weights is minimal in each net in the class considered, and (3) the realization procedure incorporates maximum protection against disconnection of the net due to either branch or node erasure.

INTRODUCTION

Nets in which both the branches and nodes are finitely weighted will be called weighted nets, in order to distinguish them from branch-weighted nets, in which only the branches are weighted, and node-weighted nets, in which only the nodes are weighted. The class of weighted nets considered in this paper is that in which the node weights determine the maximum flows. Hence, the terminal capacity matrix of any net in this class must satisfy the necessary and sufficient conditions for realizability as a purely node-weighted net; vis-a-vis Tapia and Myers.[1] Realization proceeds directly from the terminal capacity matrix. First, a node-weighted realization is derived; and then the minimum individual branch weights which are both necessary and sufficient for maximum flow are determined according to the topology and node weighting.

Previously published realizations of node-weighted nets suffer from the disadvantage that they do not generally provide maximum protection against disconnection of the net due to either branch or node erasure.[2,3,4] In the present paper, such protection is incorporated into the synthesis; that is, the resulting nets are "min-outage" realizations.

Tapia[2] has given an algorithm for determining a sufficient set of branch weights. Each weight, so determined, is a minimum in the sense that it cannot be reduced without reducing the terminal capacity between one or more pairs of nodes in the net below that specified. Unfortunately, however, the algorithm does not always guarantee that the sum of all branch weights in the net is a minimum. Yet it is obviously desirable to minimize this sum.

This problem of finding a set of branch weights which are both necessary and sufficient, and hence minimal, is resolved by utilizing the Capacity-Distance Product Sum property of a certain class of weighted nets; which evolves as follows.

CENTRAL DEFINITIONS AND THEOREMS

It is assumed throughout that if n_i, n_j are any two distinct nodes in a weighted or node-weighted net G, with w_i, w_j their respective weights, then $i<j$ implies $w_i \leq w_j$. Further, G contains neither any branches in parallel nor any self-loops; and G is necessarily finite, connected, and nonoriented.

Definition 1. The quantity $|i-j|$ is called the distance d_{ij} between nodes n_i, n_j in G.

Definition 2. If nodes n_i, n_j of weights w_i, w_j, respectively, are adjacent in G, and w_{ij} is the weight of the branch (n_i, n_j) connecting them, then the product of d_{ij} and w_{ij} is called the capacity-distance product k_{ij} of branch (n_i, n_j). The sum of the capacity-distance products of all branches in G is called the capacity-distance product sum of G.

Definition 3. Any two distinct nets which have the same value of K are said to be capacity-distance invariant with respect to K. A set of nets is said to be capacity-distance invariant if each distinct pair of nets in the set is invariant with respect to the same value of K.

Next, it is convenient to quote Tapia's generalization of the max-flow min-cut theorem:[2]

Theorem 1. The terminal capacity between any two distinct nodes n_i, n_j in a finite, weighted net G is given exactly by

$$t_{ij} = \min (w_i, W_{\min}(R_{ij}), w_j) \tag{1}$$

where $W_{\min}(R_{ij})$ is the minimum of the cutweights of all mixed cut-sets R_{ij}[2,5] which separate n_i, n_j in G and w_i, w_j are the respective weights of n_i, n_j.

Whence, a restricted class of weighted nets:

Definition 4. G is said to be a uniform net if

$$t_{ij} = \min(w_i, w_j) \tag{2}$$

for all distinct pairs of nodes n_i, n_j in G.

Further, a subclass of uniform nets:

Definition 5. Given a node-weighted open path (linear tree) P with terminal nodes n_i, n_j, and with n_m a node of largest weight on P. If each of the sequences of node weights from n_i to n_m, inclusive, and from n_j to n_m, inclusive, is monotonically nondecreasing along P, then P is said to be concave.

If each of the said sequences is strictly monotone and increasing, then P is strictly concave.

Definition 6. A node-weighted tree T is said to be (strictly) concave if the path between any pair of nodes in T is (strictly) concave.

Lemma 1. A concave node-weighted tree is a uniform net.

Proof: follows immediately from the definitions.

Definition 7. If a weighted net G contains at least one concave tree T, then G is said to be concave.

Whence the useful theorem:

Theorem 2. A concave net is also uniform.

Proof. The terminal capacity t_{ij} between any pair of nodes n_i, n_j in a concave tree T is given precisely by (2). Since a concave net contains at least one such tree T, the theorem follows.

It is important to note that while a concave net is necessarily uniform, the converse is not generally true. For example, the net in Fig. 1 is uniform but is not concave; though it should be noted that a reduction in the weight of any branch in this net, below that shown, will reduce at least one terminal capacity in the net. By contrast, both nets in Fig. 2 are concave and therefore also uniform.

From here on, attention is confined to an even more restricted class of nets, thus:

Definition 8. A weighted net with n nodes n_i, $1 \leq i \leq n$, such that the weight w_i of node n_i is

$$w_i = i \tag{3}$$

is called a natural net.

This restriction is a convenience, rather than a necessity. The theorems to follow are, in general, readily modified for other classes of nets; the two central theorems in this paper being theorems 3 and 4.

Theorem 3. The capacity-distance product sum K of a finite, weighted, concave, natural net G cannot be less than $\frac{1}{2}n(n-1)$, where n is the number of nodes in G.

Proof. By hypothesis, the weight of any node n_i in G is $w_i = i$. Let $w_{i,j}$, $i<j$ assumed, be the weight of branch (n_i, n_j) in G. Then

$$K = \sum_{k=1}^{n-1} k \sum_{j=1}^{n-k} w_{j,j+k} \tag{4}$$

Algebraic manipulation of (4) yields

$$k = \sum_{p=2}^{n} C_p \tag{5}$$

where

$$C_p = \sum_{j=p}^{n} \sum_{k=1}^{j-p} w_{k,j} \tag{6}$$

C_p in (6) is recognized as the weight of a min-cut branch cut-set which separates the set of nodes n_i, $i \leqq k$, from the set n_j, $j > k$, in G. By theorem 1, it is therefore necessary that

$$C_p \geqq k \tag{7}$$

So that, from (5) and (7),

$$K \geqq \sum_{k=1}^{n-1} k = \tfrac{1}{2}n(n-1) \tag{8}$$

Which completes the proof.

Note that $\frac{1}{2}n(n-1)$ in (8) is the number of branches in a complete net with n nodes. Indeed,

Theorem 4. A complete, finite, weighted, concave, natural net G with n nodes can always be found such that the capacity-distance product sum K of G has the minimum possible value $K = \frac{1}{2}n(n-1)$.

Proof. The weight of any node n_i in G is $w_i = i$. Let $w_{i,j}$, $i<j$ assumed, be the weight of branch (n_i, n_j) in G. Consider the set C_i of all branches incident at node n_i. By theorem 1 and the hypotheses, the sum of weights of the branches in C_i cannot be less than w_i whenever $1 \leqq i \leqq (n-1)$; for each such C_i constitutes a cut-set which isolates n_i in G. Further, these (n-1) of the C_i are linearly independent, this being the rank of the incidence matrix of G. Suppose that each said sum is exactly w_i (=i). Whence,

$$i = \sum_j w_{j,i} + \sum_k w_{i,k} \tag{9}$$

for all i,j,k in the range $0 \leqq j < i < k \leqq n$; with the understanding that $w_{0,i} = 0$. Let, in (9),

$$\left(i - \sum_j w_{j,i}\right) = W_i \tag{10}$$

One set of solutions for the branch weights $w_{i,k}$ in (9) is then

$$w_{i,k} = \frac{1}{2^{k-i}} W_i \quad \text{for } 1 \leqq i < k \leqq (n-2)$$

$$w_{i,n=1} = w_{i,n-2} \tag{11}$$

Notice that the exponent (k-i) in (11) is the distance between nodes n_i, n_k in G. The capacity-distance product sum K of G is

$$K = \sum_{k=1}^{n-1} k \sum_{j=1}^{n-k} w_{j,j+k} \tag{12}$$

Inserting the branch weights (11) into (12) yields, after some algebraic manipulation using (9),

$$K = \sum_{i=1}^{n-1} i = \tfrac{1}{2}n(n-1) \tag{13}$$

Thus, the net G whose branch weights are given by (11) validates the theorem.

MINIMAL MAX-FLOW MIN-OUTAGE SYNTHESIS

Synthesis of natural nets from a specified terminal capacity matrix can now proceed. Each synthesized net is to exhibit max-flow, is to be minimal, and also is to be a "min-outage" realization. Max-flow is automatically ensured in that the net necessarily realizes the given matrix T. Minimality and min-outage are defined in the following sequence of developments. First, theorems 5 and 6, which follow from theorems 3 and 4.

Theorem 5. In a complete, weighted net G with n nodes n_i, $1 \leq i \leq n$, whose respective weights w_i are natural numbers with $w_i < w_j$ whenever $i<j$, the capacity-distance product K of G is bounded according to

$$K \geq \sum_{i=1}^{n-1} w_i \tag{14}$$

(Notice that (14) reduces to (8) in the event that G is a natural net; i.e., in the event that $w_i = i$ for all i.)

Theorem 6. A complete, weighted net G with n nodes n_i, $1 \leq i \leq n$, whose respective weights w_i are finite, natural numbers with $w_i < w_j$ whenever $i<j$, can always be found such that the capacity-distance product sum K of G is the minimum possible value

$$K = \sum_{i=1}^{n-1} w_i \tag{15}$$

Proofs of theorems 5 and 6 follow the patterns of those for theorems 3 and 4, and accordingly are not given here. Suffice it to state that a sufficient set of branch weights for the net of theorem 6 is given by equation (11), as for the net of theorem 4, except that the quantity W_i therein is now given by

$$W_i = (w_i - \sum_j w_{j,i}) \tag{16}$$

instead of by (10); and, of course, (16) reduces to (10) in the event that G is a natural net, when $w_i=1$.

Consider next some reasoning leading to a suitable definition of a "min-outage" net G. By this is meant, or intended to mean, that maximum protection against disconnection due to either branch or node erasure is incorporated in G. In the first place, therefore, each node cut-set C_{ij} of minimum cut-weight in G must contain the maximum permissible number of nodes. In realizing G from a given terminal capacity matrix T, the maximum permissible number of nodes in each such C_{ij} is easily determined, with identification of these nodes, as described implicitly in the necessary and sufficient conditions on T given by Tapia and Myers.[1,2] In the second place, G must also contain the maximum permissible number of branches. In this respect, however, Tapia and Myers[1,2] have also shown that any branches connecting the nodes in any min-cut node cut-set C_{ij} in G are redundant; that is, that such branches serve no useful purpose, and thus might just as well be omitted from G. Whence:

Definition 9. Given G, a weighted net which realizes a given terminal capacity matrix T, with C_{ij} any node cut-set of minimum cut-weight in G or in any sub-net of G. Let G_i, G_j be any two sub-nets separated by C_{ij} in G. If then (a) no two of the nodes in C_{ij} are adjacent in G, (b) C_{ij} contains the maximum permissible number of nodes, (c) G_i, G_j are each complete and (d) each node of G_i and of G_j is adjacent to each node of C_{ij} in G, G is said to be a min-outage realization of T.

Minimality is readily defined, thus:

Definition 10. A weighted net G whose capacity-distance product sum K has the minimum possible value is said to be minimal.

For example,

Lemma 2. The complete nets of theorems 4 and 6 are minimal.

Proof of lemma 2 is immediate from definition 10 and the said theorems. Of more importance, and immediate relevance, is the following theorem.

Theorem 7. Let H_i be the sub-net of the min-outage net G of definition 9 comprising the sub-net G_i, the nodes C_{ij}, and all branches connecting C_{ij} and G_i. (Note that H_i is thus necessarily connected; and in the event that C_{ij} contains more than one node, H_i is also incomplete.) Then a set of finite branch weights can always be found such that, with these weights assigned to the branches in H_i, H_i is minimal.

Proof. It is left for the reader to show, after the pattern of the proof of theorem 6, that the set of branch weights obtained in the proof of theorem 6, as given by equation (11) in conjunction with (16), is sufficient to validate the theorem; wherein the branch weights in (11) corresponding to the distinct node-pairs in C_{ij} are each put to zero, since by definition 9 these are the only non-existent branches in H_i. It is also left for the reader to show, as a perfectly straightforward matter, that:

Theorem 8. The sub-net H_i of theorem 7 is concave.

(Author's note: the straightforwardness of the proofs of theorems 7 and 8 is a matter of recognizing that, by the definition of C_{ij}, any node of H_i which is of larger weight than any other node of H_i cannot be one of the nodes of C_{ij}.)

Clearly, therefore, minimal min-outage realization of a terminal capacity matrix T of a weighted, natural net G in which the max-flows are determined by the node weights in G, can be accomplished if a necessary and sufficient number of submatrices of T can be identified such that each of these submatrices is realizable by a sub-net of the type H_i in theorem 7. By theorems 1 and 8, it then remains only to connect every two such distinct sub-nets H_i,H_j by superposing their common cut-set nodes C_{ij}.

This mechanization can, indeed, be achieved; as follows.

REALIZATION TECHNIQUE

After Tapia and Myers,[1,2] the specified terminal capacity matrix is assumed to be in the bracketed principal partitioning form with a uniform bracket U, and principal submatrices A,B,C; as shown in Fig. 3. It is assumed that the reader is familiar with the properties of U,A,B and C. First, the following lemma follows directly from theorems 2, 7, 8 and the properties of a uniform bracket:

Lemma 3. The terminal capacity matrix T_i of the sub-net H_i of theorem 7 can always be put into a bracketed principal partitioning form comprising only a uniform bracket U_i.

Next, the ensuing lemmas and theorems; which again, because space limitations preclude otherwise, are given without proofs. The proofs are, however, fairly straightforward in the light of the presentations up to this point.

Lemma 4. Suppose K is any A- or B-submatrix in T, with C the submatrix which separates A from B in T. If c is the constant element in C, then

$$c = \sum_{k=1}^{m} u_k \tag{17}$$

where the u_k in (17) are uniform bracket elements in the same column of T as c.[1,2] Further suppose that m in (17) is the maximum possible value it can assume, as determined by T; and that U_m is the uniform sub-bracket in T defined by the m elements u_k in (17). Let U_k be the submatrix in T which consists of and only of the matrix K bracketed by a uniform bracket of order m whose elements are the m elements u_k in (17). Then U_k is itself a uniform bracket.

Theorem 9. If Q is the number of distinct A- and B-submatrices in T, then T is the union of exactly Q distinct uniform brackets U_k.

Lemma 5. Each of the Q distinct uniform brackets U_k in T, as theorem 9, is realizable as a sub-net G_k of the type H_i of theorem 7; wherein the min-weight node cutset nodes C_{kj} common to any two such distinct sub-nets G_k, G_j can always be identified.

Theorem 10. Given a terminal capacity matrix T which is realizable as that of a natural, weighted net, the net G comprising the Q sub-nets G_k of lemma 5, connected by superposing the cut-set nodes C_{kj} which are common to each pair of such distinct sub-nets G_k, G_j, is a minimal, min-outage realization of T.

Mechanization of the synthesis follows immediately as in the following didactic example.

DIDACTIC EXAMPLE

The terminal capacity T in Fig. 4(a) is realizable as a weighted, natural net G. Using previous notation, Q=3 for this matrix. The three corresponding uniform submatrices U_1, U_2, U_3 are shown in Fig. 4(b), and their respective realizations G_1, G_2, G_3 in Fig. 4(c). Whence the resulting minimal, min-outage realization G of T in Fig. 4(d), obtained by superposing the appropriate cut-set nodes of G_1, G_2, G_3.

Note that minimal, min-outage realizations of a given T are not necessarily unique. For, any node n_i in a uniform sub-net G_i which has a node weight w_i less than that of the cut-weight c_{ij} which separates G_i from G_j may be erased from G_i and inserted in G_j. For example, node c may be erased from G_1 in Fig. 4(c) and inserted in G_2; wherein, of course, the branch weights for G_1 and G_1 need to be re-computed. Notice also that the weights of nodes d,e,f in G, Fig. 4(d), may be assigned any values $w_d \geqq 3$, $w_e \geqq 3$, $w_f \geqq 1$, respectively, instead of those shown. The latter values were used merely to emphasize the fact that G is, to all intents and purposes, a natural net.

REFERENCES

1. B. R. Myers and M. A. Tapia, Flow through node-weighted networks, Proc. Third Colloquium on Microwave Communication, Budapest, Hungary (1966) (in press).

2. M. A. Tapia, Analysis and synthesis of weighted communication networks, Tech. Rep. No. EE-665, University of Notre Dame (1966).

3. S. S. Yau, Synthesis of radio=communication nets, IRE Trans. on Circuit Theory, CT-9, 1 (1962).

4. B. R. Myers and M. A. Tapia, Analysis and synthesis of node-weighted networks, Proc. Conference on Electrical Network Theory, University of Newcastle upon Tyne, England (1966) (in press.)

5. S. S. Yau, A generalization of the cut-set, J. of the Franklin Institute, 273, 1 (1962).

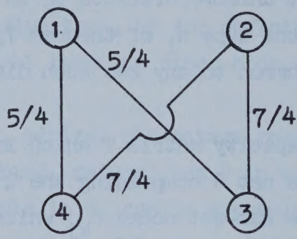

1
2
5/4
5/4
7/4
7/4
4
3

Figure 1

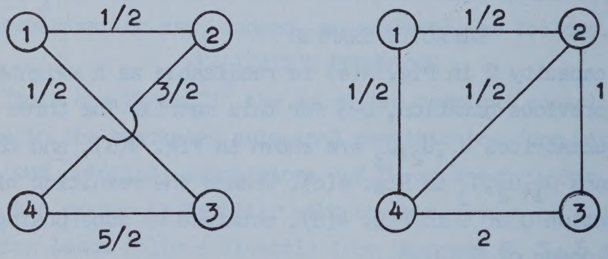

1
1/2
2
1/2
3/2
4
3
5/2
1
1/2
2
1/2
1/2
1
4
3
2

Figure 2

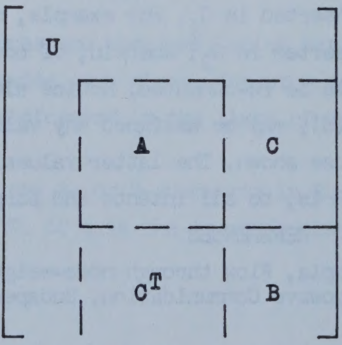


Figure 3

T

	a	b	c	d	e	f
a	-	1	1	1	1	1
b	1	-	2	2	2	1
c	1	2	-	3	3	1
d	1	2	3	-	3	1
e	1	2	3	3	-	1
f	1	1	1	1	1	-

(a)

	a	b	c	d
a	-	1	1	1
b	1	-	2	2
c	1	2	-	3
d	1	2	3	-

U_1

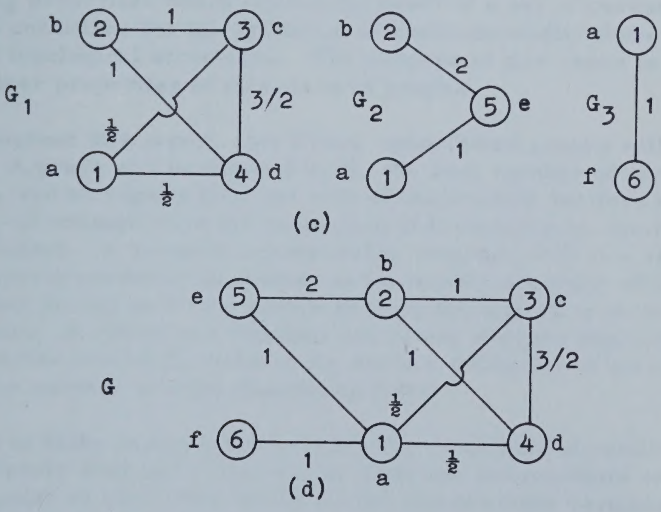
	a	b	e
a	-	1	1
b	1	-	2
e	1	2	-

U_2

	a	f
a	-	1
f	1	-

U_3

(b)



G_1
b
2
c
3
a
1
d
4
1
1
½
3/2
½
G_2
b
2
e
5
a
1
2
1
G_3
a
1
f
6
1
(c)
G
e
5
b
2
c
3
f
6
a
1
d
4
2
1
1
1
3/2
½
1
½
(d)

Figure 4

ANALYSIS AND SYNTHESIS OF NODE WEIGHTED NETWORKS FROM THE SIMULTANEOUS FLOW STANDPOINT

Enrique A. Davila

Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

ABSTRACT

This paper concerns the analysis and synthesis of node weighted networks when considering simultaneous flows between two pairs of nodes. The results obtained may in some cases be generalized for flow between "m" pairs of nodes.

The concept of a disconnecting set consisting only of nodes is introduced and employed. Necessary and sufficient conditions for a set of simultaneous flows to be feasible are found when the two pairs of terminals fall into a certain class. An example is given to show why this result cannot be made to include all cases. The concepts of redundant node and proper flow path are also introduced. Some lemmas and theorems whose application reduce the graph and the hence complexity of the problem are also presented. The problem is also given a linear programming formulation which can be generalized for k simultaneous flows, k > 2. Optimal simultaneous-flow networks are defined and some synthesis of this type of networks is done. It is shown that synthesis of a optimal simultaneous-flow tree network is unique, yielding a star graph with the weight of the center node being equal to the sum of the weights of the second and fourth highest weighted nodes.

A CLASS OF PLANAR GRAPHS CONTAINING HAMILTONIAN CIRCUITS

D. T. Tang

IBM Watson Research Center
Yorktown Heights, New York

ABSTRACT: A class of graphs called "bi-path" graphs was recently shown to have some interesting properties with which the proof of a set of necessary and sufficient conditions for the existence of multicommodity flows in such networks was established. The most important property of a bi-path graph is that each of its nonseparable subgraphs with five vertices or more contains a local Hamiltonian circuit with respect to which there exist no cross-over connections. Some additional properties of this class of graphs are presented in this paper. It is shown that, in such a graph, there exist vertices of degree two and that the total number of edges cannot exceed 2n - 3, where n is the total number of vertices in the graph. Some interesting properties of the dual of a bi-path graph are also studied. For instance, it is shown that the dual of a nonseparable bi-path graph with n = 5, as well as the graph itself, is 3-chromatic.

INTRODUCTION

In the recent study of the existence of multicommodity flows in non-oriented (undirected) networks[1], a class of graphs was shown to have some interesting properties which lead to the proof of a set of necessary and sufficient conditions for the existence of multicommodity flows in networks with such topological structures. The purpose of this paper is to present some further properties of this class of graphs.

Throughout this paper, only finite, nonoriented graphs will be considered. A graph will be denoted by G, the total number of vertices (nodes) in G by n, and an edge in G by the pair of its incident vertices in parenthesis. A graph G is nonseparable (bi-connected) if it contains no cut-vertex (point of articulation). A maximal nonseparable subgraph of G is a subgraph which is not properly contained in another nonseparable subgraph of G. A Hamiltonian circuit in G is a circuit passing through all n vertices of G exactly once. A cutset is a minimal collection of edges separating two disjoint vertex sets of G, and a cycle set is a collection of all edges in a circuit or a number of edge-disjoint circuits.

A set of paths in a graph G between the same pair of vertices are said to be "properly disjoint" if there is at least one intermediate vertex (distinct from the pair) on each path, and if all the intermediate vertices on these paths are distinct.

Of present interest are graphs with the following property: Between any pair of vertices of G, there do not exist three or more properly disjoint paths. For brevity, such graphs are referred to as "bi-path" graphs. It follows that if a graph is not bi-path, it must contain at least three properly disjoint paths between a certain pair of vertices, implying $n \geq 5$.

SOME IMPORTANT PROPERTIES OF BI-PATH GRAPHS

The following four theorems have been established.[1]

Theorem 1: A bi-path graph is planar.

Theorem 2: A graph G is bi-path if and only if every maximal nonseparable subgraph of G is bi-path.

Theorem 3: A nonseparable graph G with $n \geq 5$ is bi-path if and only if it contains a Hamiltonian circuit with respect to which there exist no cross-over connections. (That is, if edges (a, b) and (c, d) belong to G, then going from a to b by means of the Hamiltonian circuit, one either passes through both or none of vertices c and d).

Theorem 4: A nonseparable bi-path graph with $n \geq 5$ contains a unique Hamiltonian circuit.

Thus a bi-path graph is characterized by the property that each of its maximal nonseparable subgraphs contains a local Hamiltonian circuit with respect to which there exist no cross-over connections, except the following two cases: The first exception occurs when a maximal nonseparable subgraph contains exactly two vertices with only one edge connecting them. The second exception occurs when a maximal nonseparable subgraph contains exactly four vertices with six edges forming a complete graph. Figure 1 shows a possible bi-path graph.

ADDITIONAL PROPERTIES

Some additional properties of bi-path graphs are now given below. Although most of them are straightforward results not difficult to see, they will be presented in the form of several theorems for the sake of clarity.

Theorem 5: In a nonseparable bi-path G without parallel (multiple) edges and with $n \geq 5$, there exist at least two vertices with degree (of incidence) equal to two.

Proof: Let C_h be the Hamiltonian circuit which exists by Theorem 3. If $C_h = G$, then all vertices are of degree two. If $C_h \neq G$, then there exists an edge (i, j) such that $(i, j) \in G$, but $(i, j) \notin C_h$. Since both i and j are on C_h, C_h provides two paths, P_α and P_β, from i to j. Let α_h and β_h be the intermediate vertex sets of P_α and P_β respectively. Since there can be no parallel edges, both α_h and β_h are not empty. Consider any vertex $a_1 \in \alpha_h$. Either a_1 is of degree two, or there exists an edge $(a_1, a_2) \notin C_h$, where $a_2 \in \alpha_h$ by theorem 3. Again, the part of P_α from a_1 to a_2 must pass through a nonempty set of intermediate vertices $\alpha_h' \subset \alpha_h$. Since the graph is finite, it is clear that the repetition of this procedure will eventually yield a vertex $a \in \alpha_h$ which is of degree two. Similarly, one can find another vertex $b \in \beta_h$ which is also of degree two.

Theorem 6: In a nonseparable bi-path graph G without parallel edges and with $n \geq 5$, there are at most $2n - 3$ edges. That is, $E_{max}(n) = 2n - 3$.
Proof: A bi-path graph is planar by Theorem 1. Let all the vertices be placed on a circle, then edges can be drawn as straight lines connecting their end vertices. Each face of a maximal bi-path graph (one which contains a maximum number of edges with respect to a fixed n) thus formed inside the circle must be bounded by exactly three edges. For if a face in G_m is bounded by four or more edges, at least one edge can be added to G_m without violating its bi-path property, contradicting the assumption that G_m is maximal. Let $x \in G_m$ be a vertex with degree two, which exists by Theorem 5. Then the removal of x and its two incident edges from G_m yields a graph G_m' with $n - 1$ vertices and $E_{max}(n) - 2$ edges. One can easily show that conditions of Theorem 3 are still satisfied by G_m' implying G_m' is also a bi-path graph, and that all the faces in G_m' remain triangular. Now there exists another vertex $x' \in G_m'$ of degree two, which can next be removed from G_m', yielding a graph G_m'' with $n - 2$ vertices and $E_{max}(n) - 4$ edges. The same operation can be repeated until $n - 3$ vertices are removed, yielding a single triangle. Therefore, $E_{max}(n) - 2(n - 3) = 3$, or, $E_{max}(n) = 2n - 3$.

Theorem 7: Let G be a nonseparable bi-path graph with $n \geq 5$, and let the vertices of G be placed on a circle with edges drawn inside the circle without intersecting one another. This uniquely defines G^* (Fig. 2), the dual of G. Let x be the vertex in G^* corresponding to the outside face of G bounded by its Hamiltonian circuit C_h. Then the removal of x and its incident edges from G^*, denoted by $G^* - x$, results in a tree (a connected circuit-free graph).

Proof: It can be shown that, since the dual of a separable planar graph is also separable, the dual of a nonseparable planar graph is also nonseparable. Therefore G^* is nonseparable. Between any pair of vertices a and b, $a, b, = x$, there exist two disjoint paths in G^*. At most one of these paths is destroyed by the removal of x from G^*, implying that $G^* - x$ is connected. Now suppose that there exists a circuit C in $G^* - x$. Then C corresponds to some cutset K in G. Since C contains no edge incident to x in G^*, C_h contains no edge of the cutset K in G. But this implies that some vertex is not on C_h, contradicting the assumption that C_h is Hamiltonian. Therefore $G^* - x$ must be circuit-free. Since $G^* - x$ is connected and circuit-free, it is a tree.

Theorem 8: A nonseparable planar graph G with $n \geq 5$ is a bi-path graph if there exists a vertex $x \in G^*$ such that the removal of x and its incident edges from G^* results in a tree.

Proof: Since the removal of x and its incident edges destroys all circuits in $G^* - x$, some bounding edge in each face of G^* must be incident to x. It follows that the face in G corresponding to x in G^* must be bounded by a Hamiltonian circuit C_h in G. Furthermore, the existence of cross-over connections in G with respect to C_h implies that G contains a subgraph homomorphic to a complete graph of four vertices. This would in turn imply that G^* also contains a subgraph homomorphic to a complete graph of four vertices. This means that no removal of a single vertex in G^* can possibly destroy all circuits in G^*, which is a contradiction. Therefore G must be bi-path by Theorem 3.

Theorem 9: A nonseparable bi-path graph with $n \geq 5$ is 3-chromatic.
Proof: Let $\bar{G}$ be a maximal bi-path graph obtained by adding edges to G until all faces are triangular. Starting with any face of $\bar{G}$, one can color the three bounding vertices with three colors. Since a neighboring face has exactly two of its bounding vertices in common with that of the first face, the third vertex gets the only color left which is distinct from that of the other two. It follows that the coloring of three vertices bounding any face of $\bar{G}$ induces a coloring (with three colors) of all other vertices in $\bar{G}$. Since the adjacency relationships of faces in $\bar{G}$ correspond to the adjacency relationships of vertices in $\bar{G}^*$, and since the induced coloring process only has to exhaust all the inside faces of $\bar{G}$, no contradictory coloring can result by Theorem 7 (which implies that the sequence of induction from one face to another is unique as the path from one point to another in a tree is unique). Therefore $\bar{G}$ is 3-chromatic, and so is G.

Since a graph is k- chromatic if and only if all of its maximal nonseparable subgraphs are k-chromatic, a bi-path graph is 3-chromatic except when at least one of its maximal nonseparable subgraphs is a complete graph of four vertices, in which case it becomes 4-chromatic.

Theorem 10: A dual of a nonseparable bi-path graph G with $n \geq 5$ is 3-chromatic.

Proof: A graph which is 2-isomorphic[2, 3] with another k-chromatic graph is clearly k-chromatic. Since any dual of G is 2-isomorphic with the dual defined in Theorem 7, it suffices to show that G^*, the dual of G thus defined, is 3-chromatic. From Theorem 7, there exists a vertex $x \in G^*$ such that its removal results in a tree. Since a tree is 2-chromatic, the vertex x can always be colored by the third color. Therefore G^* is 3-chromatic.

With Theorem 10, it can be shown that the dual of a bi-path graph with no dangling edges (or equivalently, with no vertices of degree one) is 3-chromatic except when one of the maximal nonseparable subgraphs of the bi-path graph is a complete graph of four vertices, in which case the dual graph becomes 4-chromatic.

Theorem 11: In a nonseparable bi-path graph G with $n \geq 5$, there exist two cycle sets in G such that the "exclusive or" of these two sets results in a Hamiltonian circuit.

Proof: Consider the dual graph G^*. Let the vertices be colored according to Theorem 10 with the vertices paritioned into three disjoint sets:

$S_1 = x$, S_2, and S_3. Clearly, the "exclusive or" of the cutsets separating S_2 and S_3 respectively from other vertices in G^* results in the incident cutset of x. The dual statement in G then requires that the "exclusive or" of two cycle sets C_2 and C_3 (corresponding to S_2 and S_3 respectively in G^*) results in the Hamiltonian circuit C_h (corresponding to x in G^*).

SOME CONCLUDING REMARKS

We note that Theorems 7 and 9 together characterize a nonseparable bi-path graph with five vertices or more in terms of its dual graph. Thus, these two theorems can be considered as a dual statement of Theorem 3. Theorem 9 and 10 assure us that the dual of a nonseparable bi-path graph with n = 5, as well as the graph itself, is 3-chromatic. The converse, however, is not true. Similarly, the converse of Theorem 11 is not true. Counter examples can easily be constructed to verify the above statements.

It is interesting to see how alternative proofs of some of the above theorems can be obtained. For instance, Dirac[4] showed that in a nonseparable finite graph without parallel edges and with degree of vertices at least three, there exists a subgraph homomorphic to a complete graph of four vertices. It can be shown that this leads to an alternative proof of Theorem 5. Also, one may use the fact that a graph is k-chromatic ($k \geq 3$) if any of its transformed images resulting from a sequence of elementary operations is k-chromatic ($k \geq 3$). Here an elementary operation refers to either the operation of replacing a set of series edges by a single edge (thus eliminating vertices of degree two), or the operation of replacing a set of parallel edges by a single edge. An alternative proof to Theorems 9 and 10 can be obtained by showing that a nonseparable finite bi-path graph G with $n \geq 5$, as well as its dual G^*, can be reduced to a triangle by a sequence of elementary operations.

REFERENCES

1. Tang, D. T., "Bi-path Networks and Multicommodity Flows," IEEE Transactions on Circuit Theory, Vol. CT-11, pp. 468-474, December, 1964.

2. Whitney, H., "2-isomorphic Graphs," Amer. J. Math., Vol. 55, pp. 245-254, 1933.

3. Whitney, H., "Non-separable and Planar Graphs," Trans. Amer. Math. Soc., Vol. 34, pp. 339-362, 1932.

4. Dirac, G. A., "A Property of 4-chromatic Graphs and Some Remarks on Critical Graphs," J. London Math. Soc., Vol. 27, pp. 85-92, 1952.

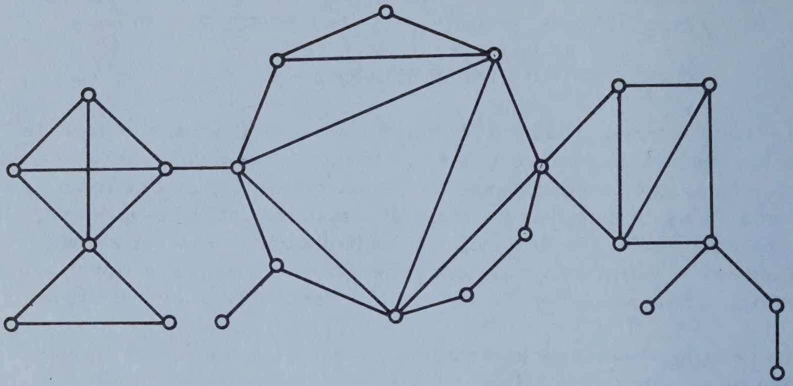


Figure 1

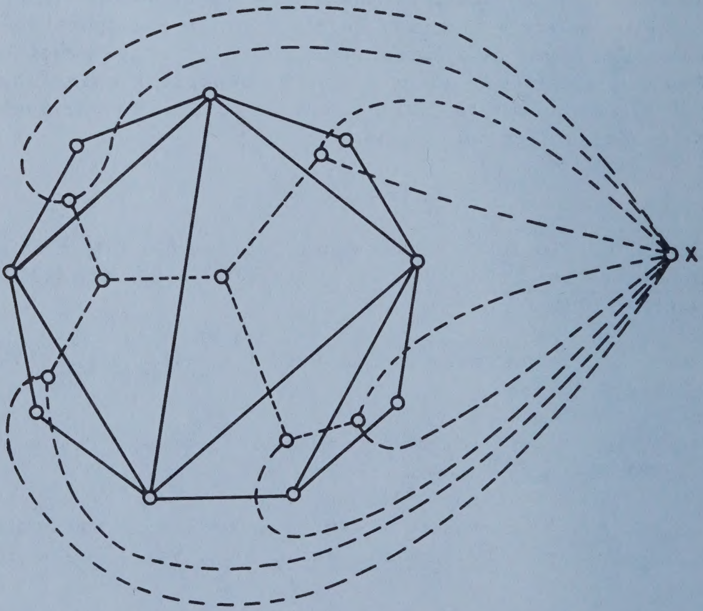


Figure 2

ADDITIONAL SIMPLE CODES FOR PLANAR GRAPHS [†]

LOUIS WEINBERG
Department of Electrical Engineering
City College of the City University of New York
New York, N. Y.

ABSTRACT

Two ordering algorithms are given for a planar triply connected graph or any plane graph. The first algorithm separately orders the meshes, branches, and nodes of a graph of connectivity ≥ 2; the second orders the star cut-sets, branches, and nodes of any connected plane graph. A set of mesh codes is formulated based on the first algorithm; a set of star cut-set codes, based on the second algorithm, is formulated in an analogous manner. Each set leads to the definition of a unique code matrix and a canonic code; each code matrix and each canonic code is unique and in one-to-one correspondence with its associated graph. The codes are dual codes; for example, a graph is self-dual if and only if the canonic mesh code is equal to the canonic star cut-set code. The codes are useful in digital-computer analysis of physical systems represented by linear graphs. They solve, for example, the isomorphism and automorphism group problems in a direct and simple manner. They are useful in the problem of squaring the rectangle, in compiling dictionaries of linear graphs, and in the solution of many other problems.

INTRODUCTION

As is well known, linear graphs are increasingly being used as models for physical systems. Such systems include pattern recognition, electric networks, the automatic retrieval of information, machine translation of languages, and all sorts of commodity-flow and information-flow systems. The representation of systems by graphs and the analysis of these systems by a digital computer require the ability to transmit the graphs from point to point within a computer, to extract information from them, and to modify and otherwise operate on them. We are concerned with a few of the problems that arise in attempting to do this for planar graphs. Specifically, in this paper we show how to encode and decode graphs; in a companion paper[1] we use these new codes to formulate some new theorems and algorithms for determining whether two graphs are isomorphic and for determining the automorphism group of a given graph.

We previously demonstrated a maze-solving method[2] for generating a canonic code for a planar graph, where by canonic we mean a code that is in one-to-one correspondence with the associated graph. In that paper we also discussed the importance of such codes. For a triply connected planar graph this code was defined as the first row of a unique matrix of 4b codes, where the codes are ordered lexicographically.

In the present paper we show a number of new methods for generating a canonic code for planar graphs. The basic schemes are simpler than those in our previous paper; more important, they make the topological structure of the graph evident almost by inspection. The methods are based on two theorems, one of which is the dual of the other. One of the theorems gives an algorithm for assigning unique orderings to the nodes, the branches, and the meshes of a plane

---

[†] Most of the research reported in this paper was supported by the National Science Foundation under Grant GK 1093.

graph; the dual second theorem orders the nodes, the branches, and the star cut-sets of a plane graph. These orderings then yield canonic codes which may be derived by use of any type of incidence matrix for the graph, for example, the node-to-node matrix, the mesh-to-mesh matrix, the mesh-to-branch or the mesh-to-node matrix, and the star cut-set-to-branch matrix. Duijvestijn's mesh code, which was discussed briefly in the previous paper[2], uses the essential information in the mesh-to-node matrix. As we shall see, the mesh-to-node matrix is not most suitable for our purposes; for example, to determine whether a graph is self-dual we would like to check a simple necessary and sufficient condition like the equality of the canonic mesh code and the canonic star cut-set code. The mesh-to-branch matrix has a dual matrix, the star cut-set-to-branch matrix, that is simpler to use in our procedures than a dual of the mesh-to-node matrix; in other words, it permits the formulation of such a condition. However, the mesh code based on the mesh-to-node matrix is useful for many purposes, it already has a large number of computer programs written for it, and since the code is not unique, it makes clear the problems that must be solved before a unique code can be found; we therefore include a brief discussion of the mesh code formulated by Duijvestijn.

Because of space limitations we assume the reader is familiar with linear graph concepts and, more specifically, with our previous paper on codes[2]; familiarity with two other papers [3, 4] on the use of the maze-solving codes to solve isomorphism and automorphism group problems is also desirable. Thus only a few definitions are given below and in the body of the paper.

The concept of a planar graph is well known: it is merely a graph that can be drawn in the plane without crossed branches. A planar graph that is so drawn is designated as a plane graph. A graph in which at most one branch joins each pair of nodes is a simple graph; a graph with more than one branch incident to the same pair of nodes, that is, with multiple or parallel branches, is a multigraph.

The branches of a plane graph G form a set of adjoining polygons which divide the plane into regions called faces; thus a face is a region bounded by a circuit where the region contains no branches or nodes of G. This particular type of circuit is termed a mesh. We define a mesh as a cyclically ordered set of branches (or nodes) that encircle the corresponding face in a clockwise sense. Dually, a star cut-set is a cyclically ordered set of all branches incident to a node, where the ordering is such that the node is encircled in a clockwise sense; alternatively, the (label of the) node at the tip-end of each branch in the star cut-set may be used in the ordered set.

It is important to note that representing a planar graph as a plane graph induces a fixed cyclic order on the branches (and nodes) of each mesh and each star cut-set. In general, a different plane representation of the same graph will yield a new set of meshes and a new set of star cut-sets. However, for a simple triply connected graph there are only two plane representations, namely, any plane representation of the graph and the mirror-image graph[3]. In other words, the set of meshes and the set of star cut-sets are unique within a reversal of the cyclic order.

In the remainder of the paper we first discuss the mesh code[5] of Duijvestijn. Then we discuss the basic ordering theorems and their proofs. Next we consider the mesh codes and show how they are made unique. This leads to a similar consideration of star cut-set codes. The discussion is concerned mainly with undirected graphs that are triply connected, but we show briefly how the tree, the series-parallel graph, and the general planar graph may be handled by the methods formulated in the paper.

THE UNIQUE-CODE PROBLEM

In solving the problem[5] of "squaring the rectangle" crucial use is made of simple planar triply connected graphs. It is required that these graphs be coded so that they may be analyzed by a digital computer. In working on this problem Duijvestijn used a mesh code and wrote programs in ALGOL that use this code for obtaining the incidence matrix and a code of the dual graph, and for generating codes of triply connected graphs containing b+1 branches from those containing b branches.

Probably the program that takes the most computer time is not mentioned in the preceding list; this is the one for solving what Duijvestijn calls the "identification" problem, namely, the determination whether two given codes correspond to the same graph; this is very difficult to solve using his mesh codes and could easily be a barrier in generating high-order squared rectangles. Duijvestijn uses a heuristic method of "characteristics" and "weights and scores" which, with the increase in order, could require an inordinate amount of computer time. With the codes formulated by the author the identification problem, because the canonic codes are in one-to-one correspondence with the graphs, no longer exists.

We explain the mesh code of Duijvestijn below.

Duijvestijn assumes the graph is drawn on a sphere so that no face and no mesh bounding the face are singled out as the infinite face and outer mesh, respectively. Thus all the meshes are counted, and if we let F be the number of faces (or meshes), Euler's formula relating the nodes, branches, and faces of a planar graph is

$$n + F - b = 2 \tag{1}$$

where n and b are the number of nodes and branches, respectively. Equivalently, we can assume that the graph is drawn in the plane as long as we agree to include the outer mesh in the code. It is of course clear that F-1 meshes of a plane graph define the graph uniquely, but it is convenient for programming purposes to code all F meshes. Similarly, though k nodes on the contour of a k-mesh prescribe the mesh, Duijvestijn also finds it convenient, as we shall see below, to repeat the first node at the end of the code for a mesh.

The code of a graph consists of the codes of its meshes; hence we must first describe the simple method for coding a mesh. We assume that all the nodes of the graph are numbered arbitrarily by the integers 1, 2, . . . , n, and we designate a node number by Y_r. We also arbitrarily number the meshes of the graph with the integers 1, 2, . . . , F, and assume, as stated previously, that the positive sense for describing a face is clockwise, so that the infinite face must be encircled counterclockwise. Though we do not require the labelling of branches for Duijvestijn's mesh code, branch labelling is desirable for a mesh-to-branch code where each mesh is defined by a cyclically ordered set of its branches; we therefore arbitrarily number the branches 1, 2, . . . , b. A simple triply connected graph with numbered nodes, meshes, and branches is shown in Fig. 1; for clarity in interpreting the number assignments, we enclose the mesh numbers in circles and the branch numbers in squares.

The boundary of a face, that is, a mesh, contains a set of nodes. A code of a mesh is obtained by moving in the positive sense around the mesh, starting with one of its nodes Y_i; in the process we encounter Y_j, Y_k, Y_ℓ, ... until we return to Y_i. The sequence $Y_iY_jY_kY_\ell \ldots Y_i$ is a code of the mesh.

A possible code of mesh 1 of the graph in Fig. 1 is 15621, but we can also use 21562, 62156, or 56215 as its code. If the mesh has k nodes (and hence k branches) there are k different ordered sequences that can be used for the code of the mesh. We can, of course, make the sequence unique by specifying a first node (or a first branch). This is done automatically in our procedure that is based on Algorithm 1 to be discussed in the next section.

A code of a graph is the sequence of codes of all its meshes separated by zeros; the meshes appear in the code in natural order. At the end of the code for the meshes two more zeros are added. Hence the code of the graph may be considered as a vector Y; since each branch occurs in two meshes, since the number of dimensions for each mesh is one more than the number of branches, and since there are F+1 zeros, it follows that the dimension of the vector is $2(b+F)+1$.

A code for the graph of Fig. 1 is

$$Y = 621560632606536054350145101234100 \tag{2}$$

It should be clear that this scheme can yield many codes that represent the same graph. There are three basic ways by which a different code is generated. First we note that a different code can be obtained by numbering the nodes in a different way; since there are n nodes, there are n! permutations of node numbering. Secondly, the chosen codes of the meshes may be permuted in the code for the graph, that is, we can also renumber the meshes; there are F ! ways to number the meshes. Thirdly, even with given node and mesh numberings, many different codes may still be obtained because of the possible cyclic permutations of the code for a mesh; if k_i represents the length of mesh i, then there are $k_1k_2k_3 \ldots k_F$ different codes introduced by these cyclic permutations. Thus the number of codes that represent the same graph is very large indeed, increasing at least as the product of the factorial of the number of nodes and the factorial of the number of meshes.

If the code is to be made unique, these three difficulties must be removed; our ordering algorithm and coding scheme accomplish this.

Now we turn to another mesh code. This code may be based on the essential information in a mesh-to-branch incidence matrix. Thus the code for mesh 1 in Fig. 1 is given by 6951 or by any code obtained by repeated cyclic permutations of this code. Here, it should be noted, we do not find it necessary to repeat the first branch at the end of the code. Using zeros and double zeros as before, we then have as a code for the graph in Fig. 1

$$W = 695105,10,4098,10,073802760143200 \tag{3}$$

Since in coding all F meshes each branch appears twice and there are F+1 zeros, the dimension of the W-code is $2b + F + 1$.

We can also write a code based on the star cut-sets, where the branches are indicated either by their tip-node labels or by their branch labels. It is convenient to use the dual elements to those in the W-code, that is, meshes and branches are replaced by star cut-sets and branches; thus we use branch labels in the code for a star cut-set. If we consider that a node number also represents the number of the star cut-set at the corresponding node, use of the same procedure on the star cut-sets yields

$$S = 26101540834, 10, 07230678909, 10, 500 \qquad (4)$$

The S-code, it is clear, has a dimension equal to $2b+n+1$.

The S-code and the W-code, as presently formulated, obviously suffer from the same disadvantages as the Y-code. Each shall be made unique by use of our ordering algorithms.

Finally, it should be mentioned that the ordering algorithms also permit the Y-codes to be used in a way that gives an equivalent statement to the previously mentioned simple condition for determining whether a graph is self-dual, namely, the canonic mesh code for the graph is equal to the canonic mesh code for the dual; it is required, however, that we use the S-code for the given graph in an intermediate step. We assume that the Y-codes and S-codes have been generated for a given graph to which the ordering algorithms have been applied. We then consider an S-code as a mesh code, that is, as a W-code, and number the nodes (in accordance with the numbered branches and meshes), just as they would be numbered by application of the mesh ordering algorithm. This is the ordered dual graph so that now a Y-code can be written for this graph. The method is used on all the codes in the star cut-set matrix to generate a mesh-code matrix of Y-codes for the dual graph. It should become clear later that the condition for self-duality applies to the Y-codes generated in this manner. It does not apply, however, to the Y-codes and star cut-set codes, even when we define a new star cut-set code that uses tip-node labels for writing the codes.

ORDERING THEOREMS FOR PLANE GRAPHS

For the discussion below let G be a simple planar triply connected graph; as will become apparent, Theorem 1 also applies to any plane graph of connectivity ≥ 2, while Theorem 2 applies to any connected plane graph. The basic ordering theorem for meshes is given by

Theorem 1. Designation of a branch of G, with a specification of its first node and its second node, and designation of a right mesh containing this branch fixes, by use of Algorithm 1 discussed below, the first branch of the ordered set of each mesh, orders the branches uniquely, orders the nodes uniquely, and orders the meshes uniquely.

The dual ordering theorem for star cut-sets is

Theorem 2. Designation of a branch of G, with a specification of its first node and its second node, and choice of the next clockwise branch to the designated branch in the star cut-set at the first node fixes, by use of Algorithm 2 discussed below, the first branch of the ordered set of branches in each star cut-set, orders the branches uniquely, orders the nodes uniquely, and orders the star cut-sets uniquely.

If the graph is a plane graph, then the specification of the right mesh in the mesh ordering algorithm and the specification of the next clockwise branch in the dual algorithm are not necessary, since they are given by the plane representation; the theorems also apply to these graphs. Moreover, the graph need only be biconnected for the mesh ordering algorithm and only connected for the dual algorithm. For the abstract triply connected graph the specifications are required in order to fix uniquely one of the two possible plane representations.

The mesh ordering procedure is given by

Algorithm 1. Let the designated branch of G be labelled 1, its first and second nodes be labelled 1 and 2, respectively, and the right mesh be labelled 1 and designated as M_1. Now consider the plane representation of G for which the specified right mesh is the right mesh for the designated branch. Using natural order we now number the branches of M_1 and number the nodes of M_1 in the defined sense for a mesh, that is, in clockwise order. We specify branch 1 as the first branch of M_1 and thus can write the unique ordered set for M_1 in terms of its branches. We then consider the branches of M_1 in order and choose the earliest branch that is part of another mesh that has not yet been labelled. By our construction this is branch 1; we label the new mesh as mesh 2, designated by M_2, and label its unlabelled nodes and its unlabelled branches by the integers in natural order (and proceeding in the clockwise sense), in each case using the first integer that has not yet been used. We consider the chosen earliest branch of M_1 that is a part of M_2 as the first branch of M_2 and write the unique ordered set of branches for M_2. Thus, if M_1 is a k-mesh, the first unlabelled branch and node of M_2 are labelled with the (k+1)st integer. We then determine the next earliest branch of M_1 that is part of a mesh that has not yet been labelled, and repeat the process, thus labelling mesh 3 and its nodes and branches and specifying the branch of M_1 by which M_3 has been reached as the first branch of M_3 so that its cyclic ordered set of branches becomes unique. After exhausting all such branches in M_1, we go on to M_2 and repeat the whole step. This step is repeated for each mesh in order and stops after mesh F and its branches and nodes have been labelled, and the ordered set for mesh F has been specified uniquely.

The procedure for ordering star cut-sets, nodes, and branches is given by

Algorithm 2. Let the designated branch of G be labelled 1, its first and second nodes be labelled 1 and 2, respectively, and the star cut-set at node 1 be labelled 1 and designated as C_1; C_1 has the specified next branch as the next clockwise branch to branch 1. Now consider the plane representation of G that has the star cut-set C_1, that is, with the correct clockwise order of branches at node 1. We specify branch 1 as the first branch of C_1 and using the integers in natural order we now number the branches and the nodes of C_1 in clockwise order. In numbering the nodes of a star cut-set we number the unlabelled tip-nodes of the branches in order; thus for C_1 branch 1 has its tip-node already labelled 2 and the tip-node of the next clockwise branch is labelled 3. We then consider the branches of C_1 inorder and choose the earliest branch that is part of

a star cut-set that has not yet been labelled. This earliest branch is branch 1; we label the new star cut-set as 2, designated by C_2, and label its unlabelled nodes and unlabelled branches in clockwise order, in each case using the integers in natural order and starting with the first number that has not yet been used; thus C_2 is the star cut-set at node 2. We consider the chosen earliest branch of C_1 that is a part of C_2 as the first branch of C_2 and write the unique ordered set of branches for C_2. We then determine the next earliest branch of C_1 that is part of a star cut-set that has not yet been labelled, and repeat the process, thus labelling C_3 and its nodes and branches, and specifying the branch of C_1 by which C_3 has been reached as the first branch of C_3 so that its ordered set of branches is unique. After exhausting all such branches in C_1, we go on to C_2, and repeat the whole step. This step is repeated for each star cut-set in order and it stops after star cut-set C_n and its branches and nodes have been labelled and the ordered set for C_n has been specified uniquely.

It should be noted that though there is a correspondence between the labelling of meshes and branches in Algorithm 1 and the respective labelling of star cut-sets and branches in Algorithm 2, there is no correspondence in the labelling of nodes in the algorithms. This accounts for the previously mentioned advantages in using branch labels rather than node labels.

Furthermore, it is pointed out that in Algorithm 2 we automatically find a tree path from the node of star cut-set 1 to the node of each of the other star cut-sets; the branches of this tree are given by the first branches of C_2, C_3, ..., C_n. Analogously, in Algorithm 1 we determine a path across the branches of a co-tree from mesh 1 to every other mesh, where the branches of the co-tree are the first branches of M_2, M_3, ..., M_F. This accounts for our use of the phrase reach a mesh or a star cut-set.

We now prove Theorem 1; an analogous proof holds for Theorem 2.

Proof. The proof proceeds by induction.

We first observe that a plane representation is specified by the designated branch and direction and its right mesh. Thus the set of meshes is unique, where each mesh is a unique set of branches with a specified cyclic order. Mesh 1 is specified uniquely as the designated right mesh and specification of the designated branch as its first branch makes its cyclically ordered set a unique ordered set. Starting with the first branch and first node the previously unlabelled branches and nodes of M_1 are ordered uniquely by the integers in natural order. Using the branches of the k-mesh M_1 in order in accordance with the algorithm fixes the unique labelling of the meshes M_2 through M_{k+1}. The first branch of M_2 is uniquely specified by the algorithm as the earliest branch of M_1 that is not part of a previously labelled mesh, which in this case is the designated branch; its unlabelled branches and nodes are labelled uniquely in clockwise order by the integers in natural order, in each case using the first integer that has not yet been used. A corresponding uniqueness is achieved in the same way for meshes M_3 through M_{k+1}.

We now consider the branches in order in the ordered set for M_2 and repeat the process. The proof is then completed by induction on the ordering of the meshes.

We assume that uniqueness has been achieved for M_n and its nodes and branches, where $2 < n < F$. Suppose that a branch of M_r, where $2 \le r < n$, has been used in determining the labelling of M_n. The next earliest branch that is part of a mesh that has not yet been labelled occurs in M_r or it does not. If it does, it is uniquely determined and thus specifies the mesh labelled n+1 and we carry out the unique ordering of its nodes and branches and specification of the ordered set for M_{n+1}. If it does not, we consider the succeeding meshes in order until this branch is determined uniquely in M_p, and thus again the specification of M_{n+1} and the orderings of its nodes and branches are unique. Since the graph is biconnected, a path exists across the branches of a co-tree from M_1 to all the other meshes so that M_p can be found where $2 < p < n+1$. Hence the orderings are unique and the proof is complete.

An example of the application of Algorithm 1 is given in Fig. 2a and of Algorithm 2 in Fig. 2b, where we have used the plane graph in Fig. 1 with the branch (1, 2) as its designated directed branch; again for clarity the branch numbers are in squares and the mesh and star cut-set numbers are in circles. We note that star cut-set C_p is at node p, that is, the node number and the corresponding star cut-set number are identical so that one set of numbers is actually superfluous. Since the set of meshes in Fig. 2a is given by

$$\begin{array}{ll} M_1 = 1234 & M_4 = 39,10 \\ M_2 = 1567 & M_5 = 4,10,5 \\ M_3 = 278 & M_6 = 698 \end{array} \tag{5}$$

the mesh code is

$$W = 1234015670278039,10,04,10,5069800 \tag{6}$$

Similarly, the set of star cut-sets in Fig. 2b is

$$\begin{array}{ll} C_1 = 123 & C_4 = 3789 \\ C_2 = 145 & C_5 = 49,10 \\ C_3 = 267 & C_6 = 5,10,86 \end{array} \tag{7}$$

and the corresponding star cut-set code for the graph is

$$S = 1230145026703789049,10,05,10,8600 \tag{8}$$

It is clear that the unique orderings have removed the three difficulties that led to a multiplicity of codes for a given graph. Thus for the specified directed branch and right mesh, the mesh code is unique; analogously, for the specified directed branch and star cut-set at the first node, the star cut-set code is unique.

Thus for Algorithm 1 the choice of the first branch, its direction, and its right mesh fixes the numberings and hence the mesh code; an analogous statement holds for Algorithm 2 and the star cut-set code.

Now consider Algorithm 1 and the mesh code; a corresponding discussion holds for the dual algorithm and the star cut-set code. We now ask in how many ways a directed branch can be chosen. For a plane graph that is triply connected each branch with its right mesh may be chosen in either of its two directions; thus there are 2b choices. In the mirror-image graph, for a specified branch and direction the right and left meshes change roles; there are therefore an additional 2b choices of a directed branch and its right mesh. Thus there is a total of 4b choices, so that applying Algorithm 1 leads to 4b different mesh, branch, and node numberings. Each of these choices yields a mesh code that is in one-to-one correspondence with the graph for this choice of directed branch and right mesh. We thus obtain a total of 4b mesh codes.

In a dual manner we obtain a total of 4b star cut-set codes for a planar triply connected graph.

Each of the 4b mesh codes can be considered as a vector. These vectors are now arranged lexicographically (or more briefly, lexically), that is, like the words in a dictionary. For these codes, whose components are the integers $0, 1, 2, \ldots, b$, this is equivalent to arranging the codes as numbers ordered in increasing magnitude, where if $b \geq 10$, the code number is understood to be to the base b+1. The ordered codes are considered as the rows of a mesh code matrix H, whose dimension is thus $4b \times (2b + F + 1)$. This mesh code matrix H is unique; in fact, it should be clear that there exists a one-to-one correspondence between a planar triply connected graph G and its mesh code matrix H. The first row of H may be defined as the canonic mesh code, or where no confusion results, merely as the mesh code, for graph G. There is a one-to-one correspondence between G and its canonic mesh code.

A corresponding discussion holds for the star cut-set code matrix K and its canonic code. It should now be clear that if a graph is self-dual, its canonic mesh code is equal to its canonic star cut-set code, and conversely.

For a planar graph whose connectivity < 3 there are in general more than two plane representations[2]. For any plane representation of a graph with connectivity 2 we can define 2b mesh codes and for any connected plane graph we can define 2b star cut-set codes by the above procedures. Since the tree and the series-parallel graph have canonic plane representations, the set of 2b codes can be defined uniquely for each of them. We can thus use these 2b mesh (star cut-set) codes, lexically ordered, as the rows of a unique mesh code (star cut-set code) matrix, which is in one-to-one correspondence with the graph. The first code, defined as the canonic code, is also in one-to-one correspondence with its associated plane graph.

For the general planar graph the number of plane representations, that is, the number of distinct topological maps, has been given by MacLane[6]. This characterization may be used for specifying a unique code matrix and a canonic code for a general planar graph.

CONCLUSION

Two different sets of codes have been specified for planar graphs that are triply connected and for the tree and the series-parallel graph. These codes can also characterize the general planar graph. Each set of codes defines a unique code matrix and a canonic code, each of which is in one-to-one correspondence with its associated graph. Like the previous set of codes formulated by the author[2] the codes are useful in many computer applications of linear graphs; these codes, moreover, have the desirable property of making the topological features of the graph evident almost by inspection. In a companion paper[1] we show that the codes provide a direct solution of the isomorphism and automorphism group problems. They also permit the compilation of dictionaries of graphs. They should prove extremely useful in the problem of generating squared rectangles of high orders. They also are useful for search of patent files and in many other applications.

Though the discussion is given in terms of undirected graphs the codes also solve similar problems for directed graphs.

REFERENCES

1. Weinberg, L., "Alternative Algorithms for Determining Isomorphisms and Automorphism Groups of Planar Graphs," Fourth Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois, Oct. 1966, published in these Proceedings.
2. Weinberg, L., "Plane Representations and Codes for Planar Graphs," Proc. of the Third Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois, Oct. 1965.
3. Weinberg, L., "A Simple and Efficient Algorithm for Determining Isomorphism of Planar Triply Connected Graphs," IEEE Trans. Circuit Theory, vol. CT-13, no. 2, June 1966, pp. 142-148.
4. Weinberg, L., "Algorithms for Determining Automorphism Groups for Planar Graphs," Proceedings of the Third Allerton Conference on Circuit and System Theory, Univ. of Illinois, Oct. 1965.
5. Duijvestijn, A.J.W., "Electronic Computation Squared Rectangles," Philips Research Reports, vol. 17, Dec. 1962, pp. 523-613 (Reprint R 466).
6. MacLane, L., "A Structural Characterization of Planar Combinatorial Graphs," Duke Math. J., vol. 3, no. 3, Sept. 1937, pp. 460-472.

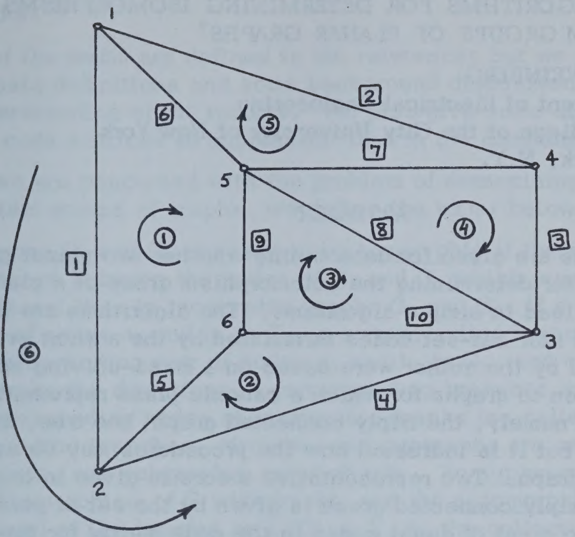
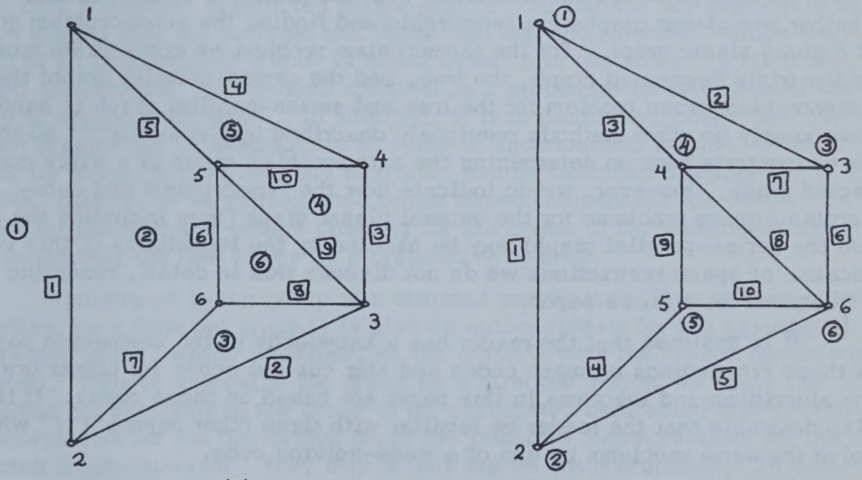

1
2
3
4
5
6
①
②
③
④
⑤
⑥
1
2
3
4
5
6
7
8
9
10

Fig. 1. A triply connected graph with labelled nodes, meshes, and branches.


1
2
3
4
5
6
①
②
③
④
⑤
⑥
1
2
3
4
5
6
7
8
9
10
1 ①
2 ②
3 ③
4 ④
5 ⑤
6 ⑥
1
2
3
4
5
6
7
8
9
10

(a) (b)

Fig. 2. Labelled graphs obtained by applying the ordering algorithms to the graph in Fig. 1.

ALTERNATIVE ALGORITHMS FOR DETERMINING ISOMORPHISMS AND AUTOMORPHISM GROUPS OF PLANAR GRAPHS†

LOUIS WEINBERG
Department of Electrical Engineering
City College of the City University of New York
New York, N.Y.

ABSTRACT

Theorems are given for determining whether two planar graphs are isomorphic and for determining the automorphism group of a planar graph; these theorems lead to simple algorithms. The algorithms are based on mesh codes and star cut-set codes formulated by the author; previous algorithms treated by the author were based on a maze-solving code. Major attention is given to graphs for which a canonic plane representation can be easily defined, namely, the triply connected graph, the tree, and the series-parallel graph, but it is indicated how the procedures may be applied to the general planar graph. Two representative theorems given in the paper are: the group of a triply connected graph is given by the set of permutations corresponding to a set of equal codes in the code matrix for the graph; two trees are isomorphic if and only if their canonic star cut-set codes are equal. Though the discussion is given in terms of undirected connected graphs, the algorithms also apply to unconnected graphs and to directed graphs.

INTRODUCTION

In this paper we are concerned with the problems of determining whether two planar graphs are isomorphic and finding the automorphism group of a given planar graph. For the isomorphism problem we concentrate mostly on the triply connected graph, the tree, and the series-parallel graph; the automorphism group problem for the tree and series-parallel graph is handled more simply by other methods previously described by the author[1,2] so that we concentrate here on determining the automorphism group of a triply connected graph. However, we do indicate how the isomorphism and automorphism group problems for the general planar graph (thus including the tree and the series-parallel graph) may be handled by the techniques of this paper. Because of space restrictions we do not discuss this in detail, reserving such a treatment for a future paper.

It is assumed that the reader has a knowledge of the companion paper[3] in these Proceedings on mesh codes and star cut-set codes for planar graphs; the algorithms and theorems in this paper are based on those codes. It is also desirable that the reader be familiar with three other papers[1,2,4] which solve the same problems by use of a maze-solving code.

Because of space restrictions we are not concerned with giving proofs of most of the theorems; these will be found in future papers. However, it is

---

[†] A major part of the research reported in this paper was supported by the National Science Foundation under Grant GK 1093.

believed that enough detail is given so that proofs can be formulated by the interested reader.

Most of the terms are defined in the references but we do present below a few basic definitions and some background discussion to aid in the immediate understanding of the results. We also give some discussion of the codes and code matrices to supplement that in the companion paper[3].

Since we are concerned with the problem of determining isomorphisms and automorphism groups of graphs, we define the terms below.

Two graphs G_1 and G_2 are said to be isomorphic if there is a one-to-one correspondence between the nodes of G_1 and G_2 which preserves the incidence relations; thus in isomorphic graphs G_1 and G_2, if a branch exists between a pair of nodes a_1 and b_1 in G_1, a corresponding branch exists between the corresponding pair of nodes a_2 and b_2 in G_2, and conversely. For directed graphs the directions of corresponding branches must also be the same. Corresponding nodes of isomorphic graphs are called similar nodes; corresponding branches, circuits, and subgraphs are called similar branches, circuits, and subgraphs, respectively. An automorphism of a graph G is an isomorphism of G with itself, and the automorphism group (or briefly, the group) of G, denoted by $\Gamma(G)$ or Γ, is the collection of all its automorphisms.

The group Γ is a permutation group acting on the set $N = 1,2,\ldots,n$, where n is the number of nodes of G. The degree and order of Γ, denoted respectively by $|N|$ and $|\Gamma|$, are the cardinalities of N and Γ, respectively.

As pointed out in a previous paper[2] in greater detail, though we consider only connected undirected planar graphs our results are applicable to planar graphs that are directed and those that are unconnected. Since the group of an unconnected graph[5] is given in terms of the groups of its connected components, all we need consider is a procedure for determining the group of a connected graph. In addition, if it is desirable, codes for connected graphs can be generalized to apply to unconnected graphs. Thus codes for trees, to consider one example, can be extended to apply to forests, that is, to an unconnected graph whose components are trees, merely by specifying a unique ordering for the codes of the component trees and separating the tree codes by a distinctive symbol.

With regard to the group of a directed graph, it is clear that an automorphism for a directed graph G is also an automorphism for the associated undirected graph, which is defined as the graph G with the branch directions removed. Thus to determine the group for a directed graph, we apply the algorithm for determining the group to its associated undirected graph. Only automorphisms for the undirected graph are possible automorphisms for the directed graph; however, they may or may not be automorphisms. To complete the test, directions are now reassigned to the branches, and the graph with an initial assignment of node numbers is now renumbered in accordance with an automorphic permutation. The incidence matrix for each number assignment is formed and the two matrices are tested for equality; the automorphism holds for the directed graph if and only if the matrices are equal.

The reader should also consult the previous papers[1,2,4] for brief discussions of how to determine whether a graph is planar, how to draw a plane graph, how to determine the connectivity of a graph, and how to draw a canonic plane representation of a tree and a series-parallel graph.

We showed in the companion paper[3] how to determine for a given planar triply connected graph a mesh code W based on the mesh ordering algorithm designated as Algorithm 1. By choice of all the possible different initial directed branches and right meshes we determine a total of 4b mesh codes; these are then arranged lexically as the rows of a matrix H. The first row of H is defined as the canonic mesh code.

It is easy to prove that H and the canonic mesh code are each in one-to-one correspondence with the given graph. If two codes are equal, they specify a one-to-one correspondence of branches such that meshes correspond to meshes. Thus by a theorem[6] previously used in the proof of the maximum order of Γ, the graphs corresponding to the two codes are isomorphic. If the codes are obtained from the same graph by use of two different combinations of designated directed branch and right mesh, the permutations of node numbering corresponding to the two codes specify an automorphism.

An analogous set of codes was defined based on a star cut-set ordering algorithm. Each star cut-set code is designated as a vector S and the star cut-set code matrix is designated by K. The first code of K is defined as the canonic star cut-set code.

The use of 4b codes in the matrix is based on the author's theorem[6] that the maximum order of the automorphism group of a planar triply connected graph is 4b. However, it is only the $|\Gamma|$ of each of the Platonic graphs that achieves this maximum; for all other planar triply connected graphs $|\Gamma|$ is smaller. This allows the use of a smaller number of codes, that is, a submatrix of each of the previously defined matrices, for determining isomorphism and the automorphism groups of planar triply connected graphs. Canonic codes can then be defined in terms of these submatrices. The requirement that a smaller number of codes need be generated is important in that computation time can be greatly decreased.

Furthermore, if we are interested only in a canonic code, that is, in checking isomorphism of graphs and not in their automorphism groups, and if a canonic plane representation exists for the class of graphs under consideration, then there is no need to use the mirror-image graph for defining an additional set of 2b codes.

We now mention some of the methods for reducing the number of codes that must be generated. No attempt is made to prove the assertions; this will be done in other papers. However, most of the assertions are provable without too much difficulty along the lines indicated above and in our previous papers.

As mentioned above, the least upper bound on $|\Gamma|$ for planar triply connected graphs is given by 4b. However, if we specify the class of such graphs by their distributions of mesh lengths and node valences, we get a smaller upper bound. We let p_k represent the number of meshes of length k, and φ_r the number of nodes of valence r, and L represent an upper bound on $|\Gamma|$ for planar triply connected graphs with specified distributions of mesh lengths and node valences. Then

$$L = \min (2k\, p_k\, ,\ 2r\, \varphi_r) \tag{1}$$

where the minimum is taken over all nonzero r and k, representing respectively the node valences and mesh lengths of this class of graphs. It is pointed out that L is not always a least upper bound; however, it is believed that the use of one more condition can fix a least upper bound on each class of graphs. This conjecture is still under study.

Thus for a given planar triply connected graph we determine L. It is possible that there is more than one value of k or of r or of k and r that yield a minimum. To achieve uniqueness we adopt a convention. If values of both k and r yield L, we choose k; furthermore, if there are two or more values of k yielding L, we choose the smallest k. If only two or more values of r yield the minimum, we choose the smallest value of r.

Suppose a value k has been determined using Eq.(1) in the above manner. For each k-mesh, we write a code for each of its branches with the k-mesh as right mesh; thus for each k-mesh we obtain k codes for one plane graph and k codes for its mirror-image graph.

We then arrange the $2kp_k$ codes in lexical order and use them as the rows of a matrix; this matrix is a submatrix of the mesh code matrix H. We designate it as Q and define its first row as the canonic code for the graph.

If a value of r has been determined, we proceed in a similar manner; we write 2r star cut-set codes using each of the branches with an initial node having the valence r. The submatrix is designated as P.

Furthermore, inspection of the proof[6] for the maximum $|\Gamma|$ shows that it is possible to use an even smaller submatrix than that specified by the $2kp_k$ or the $2r\varphi_r$ yielding the upper bound L. For example, instead of a submatrix containing $2kp_k$ codes, we can use two submatrices containing a total of $2k+p_k$ codes; since one of the codes in the p_k set is contained in the 2k codes, we need generate only $2k+(p_k-1)$ codes. This can reduce the computational labor appreciably in any implementation of these schemes.

The 2k codes are written for any one k-mesh with the specified k yielding the upper bound L; that is, we write a code for each of the k branches, using this mesh as right mesh in any plane representation and in its mirror-image graph. This forms a test set of 2k codes. Each of the other p_k codes are written for any one branch of each of the p_k k-meshes, with this k-mesh as right mesh.

The above also applies to plane graphs for which we wish to determine a canonic code. Since a canonic plane representation is known[1] for the tree and for the series-parallel graph , there is no need to consider the mirror-image graph. Thus for these graphs, if we are using a submatrix based on L, we need use only L/2 codes in defining this submatrix and its canonic code.

For the general planar graph we can also fix a canonic plane representation so that a canonic code can be defined for the purpose of testing isomorphism. We can even use these codes and techniques for determining the automorphism group, but we must expand the size of the matrix so that in general we may have to use more than 4b codes. Each distinct topological representation of the graph brought about by "rotations" of two-terminal subgraphs and permutations of subgraphs in parallel must be considered. Also in determining an upper bound corresponding to L we must consider all these representations. Then the expanded set of equal codes gives the group.

ISOMORPHISM THEOREMS

We now state a number of theorems on isomorphism of planar graphs that are based on the preceding discussion. The algorithms follow from these theorems in an obvious manner.

Let G_1 and G_2 be planar triply connected graphs for which code matrices, code submatrices, and canonic codes have been generated.

Theorem 1. Two graphs G_1 and G_2 are isomorphic if and only if their respective mesh code matrices are equal, that is,

$$H_1 = H_2 \tag{2}$$

A similar theorem holds for graphs represented by their star cut-set code matrices. We will also state the following theorems in terms of mesh codes with the understanding that the reader can generate the corresponding theorems for the star cut-set codes.

We can also use a single code for each graph or the canonic code for determining isomorphism. Hence

Theorem 2. Two graphs G_1 and G_2 are isomorphic if and only if one of the following equivalent conditions holds:

a) The canonic mesh code for G_1 equals the canonic mesh code for G_2.

b) Any one mesh code for G_1 is equal to a mesh code for G_2.

A test for isomorphism can also be based on the code submatrices Q or P and their respective canonic codes. Hence we have the two theorems:

Theorem 3. Two graphs G_1 and G_2 are isomorphic if and only if their mesh code submatrices are equal, that is ,

$$Q_1 = Q_2 \tag{3}$$

Theorem 4. Two graphs G_1 and G_2 are isomorphic if and only if one of the following equivalent conditions holds:

a) their canonic mesh codes defined with respect to their respective submatrices Q_1 and Q_2 are equal.

b) any one code of Q_1 is equal to any one code of Q_2.

We also have theorems based on the use of both the mesh code matrix and the star cut-set code matrix or their corresponding submatrices.

Theorem 5. A graph G_1 is isomorphic to its dual G_2, that is, the graph is self-dual, if and only if the mesh code matrix of G_1 is equal to the star cut-set code matrix of G_1, that is,

$$H_1 = K_1 \tag{4}$$

Theorem 6. A graph is self-dual if and only if its canonic mesh code is equal to its canonic star cut-set code.

If we wish to use the submatrices for determining self-duality, we must take account of the method of determining L by use of the mesh codes when there are "ties," that is when a value of both r and k determine a minimum. We can do this in a number of ways. If such a tie exists, we determine Q in the usual way; this is Q_1 for the given graph G_1. We then determine P by using the corresponding value of r that determines L; this is Q_2 for the dual graph G_2. We then have

Theorem 7. A graph G_1 is isomorphic to its dual G_2 if and only if

$$Q_1 = Q_2 \tag{5}$$

It is pointed out that a necessary condition for the graph to be dual is that a tie exist. Thus if Q_2 defined in the above manner does not exist, the graph is not self-dual.

A similar set of theorems can be written for planar graphs that are not triply connected. We must keep in mind that the size of the matrices may be decreased in accordance with our discussion in the Introduction; thus for a tree and a series-parallel graph we use a matrix of only 2b codes or a submatrix of L/2 codes. We have, for example,

Theorem 8. Two trees are isomorphic if and only if their star cut-set code matrices are equal.

Theorem 9. Two trees are isomorphic if and only if their canonic codes (based on their star cut-set code matrices) are equal.

The theorems also hold for series-parallel graphs. Moreover, in this case the mesh codes are also applicable.

The reader can easily formulate the corresponding theorems for the code submatrices.

For the general planar graph we have (again using a matrix of 4 b codes)

Theorem 10. Two planar graphs G_1 and G_2 are isomorphic if and only if a plane representation can be found for each graph such that their mesh code matrices are equal.

Again based on the above theorem a number of similar theorems can be formulated for the general planar graph. We do not discuss at this point the problem of how a unique plane representation can be determined; this will be considered in a subsequent paper.

THEOREMS ON AUTOMORPHISM GROUPS

In the mesh code matrix H and the star cut-set code matrix K for planar triply connected graphs there will in general be one or more sets of equal codes; these determine the elements of the automorphism group Γ.

We will again carry out the discussion in terms of the mesh code matrix H.

We choose any set of equal codes in H, say, the first set. To each of these codes there corresponds a node numbering. We specify the node numbering corresponding to the first code in this set as the original numbering; thus the permutation corresponding to this code is the identity. In addition, we list the permutations of node numbering corresponding to the other equal codes. These permutations are the elements of the automorphism group Γ. We formalize this as

Theorem 11. The automorphism group of a planar triply connected graph G is given by the permutations corresponding to the equal codes in any set of equal codes in H. If no codes are equal, the group is the identity.

We may also use the submatrices to determine the elements of the group Γ.

Theorem 12. The group of a planar triply connected graph G is given by the permutations corresponding to the equal codes in any set of equal codes in Q. If no codes are equal, the group is the identity.

We have previously mentioned that the size of the matrix must be increased for the general planar graph; we must take account of each distinct topological representation. By a distinct topological representation we mean that the plane graph has a different set of meshes. Thus for a triply connected graph we have two representations where the meshes are different in that the cyclically ordered set for each mesh is reversed in the mirror-image graph. Alternatively, we can consider that the representations are the same if their mesh sets are identical within a cyclic reversal, that is, each mesh must be the same or each mesh must undergo the reversal. However, for a general planar graph other differences may be introduced: some of the meshes may remain the same while others change, either by a cyclic reversal or by a permutation other than a cyclic permutation (the latter, of course, not being considered a different set); or some of the mesh sets may remain the same while other meshes even change their length; or all the sets may change. In each case the whole set of meshes is said to be different.

To illustrate the above we consider the series-parallel graph drawn in Fig. 1a; this graph is drawn with a different outer mesh in Figs. 1b and 1c. However, these are the same representation in that they have the same meshes within a cyclic reversal; thus the use of the 4b codes obtained from the graph in Fig. 1a and its mirror image provides for these three graphs and, in addition, for the other five graphs of the 2F = 8 graphs that can be obtained by stereographic projections of one of these representations mapped on the sphere. The set of meshes is given by $\{ABC, BDEC, DFE, FA\}$ or by a cyclic reversal of this set.

However, the plane representations given in Fig. 2 are all different in that we cannot obtain the other three by stereographic projection of the fourth mapped on the sphere. However, the graph in Fig. 2b is obtained by a trivial rotation of two branches in parallel. We call a rotation trivial in that it produces no change in the relative positions of the nodes; that is, the node numbering remains the same. This becomes apparent if for multigraphs, that is, graphs that have branches in parallel, we write the mesh sets using node labels rather than branch labels; for simple graphs, of course, we may also use node labels. This convention makes it clear that a triply connected multigraph has only two sets of meshes, differing only by a cyclic reversal and possible cyclic permutations of each mesh.

Thus if we were using a mesh matrix H or a star cut-set code matrix K we would require 4b codes for each of the graphs in Figs. 2a, 2c, and 2d, thus making a total of 12b = 72 codes.

This is an extremely large number of codes for such a simple graph. However, it is not necessary to use the expanded full matrix; there are other alternatives open to us. For each of the three different topological representations (with the meshes written in terms of node numbers) we can determine the sets of mesh lengths and observe whether there is a unique set of mesh lengths. If there is, we choose the plane representation corresponding to this set as a canonic one. If there is not a unique set, we can choose the mesh-length set for which there are the smallest number of equal sets. If this family is not unique, we can order the mesh-length sets in lowest lexical order and choose the collection of sets that has lowest lexical order. If there is a unique first set of mesh lengths, we can then treat the canonic representation and its mirror image as we do a triply connected graph; in other words, the maximum $|\Gamma|$ cannot be greater than 4b, because one-to-one mesh correspondences cannot be found for this plane representation and any of the other representations that has a different set of mesh lengths. However, in general, there will not be a unique set of mesh lengths.

An alternative always exists in that we can use the submatrices; the use of the submatrix Q decreases the required number of codes sharply. We find that for one plane representation, say, the one in Fig. 2a, since it yields the smallest L (and has a unique set of mesh lengths), $L = 2k\,p_k = 2(2)(1) = 4$. Thus we must write four codes, one for each branch in the 2-mesh in the graph of Fig. 2a, with the 2-mesh as right mesh, and the corresponding two other codes for the mirror-image graph. We need consider, in addition, only the graphs in Figs.2c and 2d. Since these graphs contain no 2-meshes, no codes need be written for them.

From the above brief discussion we can see how L should be defined for the general planar graph. The minimum is again over all r and k, but we must include all the meshes and nodes in each set of distinct plane representations that have the same set of mesh lengths.

To conclude the example, it should be clear that the four codes written for Fig.2a are equal so that $|\Gamma|$ is four. This is, of course, also clear by inspection. However, it does illustrate the procedure for a general planar graph; a complete discussion will be given in a subsequent paper.

In terms of the node numbering given in Fig.2a, the group is

$$\Gamma = \left\{(1)(2)(3)(4),\ (14)(23),\ (1)(23)(4),\ (14)(2)(3)\right\} \qquad (6)$$

CONCLUSION

In terms of mesh codes and star cut-set codes formulated by the author in a companion paper^3, theorems are given for determining whether two planar graphs are isomorphic and for determining the automorphism group of a planar graph. Algorithms follow from these theorems in a straightforward manner. The discussion is given in terms of undirected graphs, but as pointed out in previous papers, the algorithms are also applicable to directed graphs.

REFERENCES

1. Weinberg, L., "Plane Representations and Codes for Planar Graphs," Proceedings of the Third Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois, Oct. 1965.

2. Weinberg, L., "Algorithms for Determining Automorphism Groups for Planar Graphs," Proceedings of the Third Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois, Oct. 1965.

3. Weinberg, L., "Additional Simple Codes for Planar Graphs," Fourth Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois, Oct. 1966, published in these Proceedings.

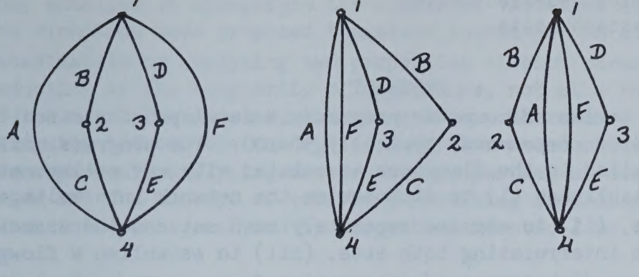
4. Weinberg, L., "A Simple and Efficient Algorithm for Determining Isomorphism of Planar Triply Connected Graphs," IEEE Trans. Circuit Theory, vol. CT-13, no.2, June 1966, pp.142-148.

5. Ore, O., Theory of Graphs, Amer. Math. Soc. Colloq. Public, Vol. XXXVIII, Prov., R.I., 1962.

6. Weinberg, L., "On the Maximum Order of the Automorphism Group of a Planar Triply Connected Graph," SIAM J.Applied Math., June 1966.

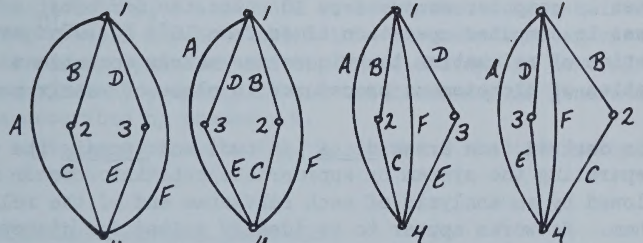
7. Weinberg, L., "Linear Graphs: Codes, New Theorems, and Algorithms for Determining their Isomorphisms and Automorphism Groups," to be published as a Univ. of Mich. Radar Laboratory Report.

8. MacLane, S., "A Structural Characterization of Planar Combinatorial Graphs," Duke Math. J., vol.3, no.3, Sept.1937, pp.460-472.


1
B
D
A
2
3
F
C
E
4
1
B
D
A
F
3
2
E
C
4
1
B
D
A
F
2
3
C
E
4

(a) (b) (c)

Fig. 1 Series-parallel graph drawn with different outer meshes.


1
B
D
A
2
3
C
E
F
4
1
A
D B
3
2
E C
F
4
1
A B
D
2
F
3
C
E
4
1
A D
B
3
F
2
E
C
4

(a) (b) (c) (d)

Fig. 2 Four different topological representations of the graph in Fig. 1.

ALGORITHMS FOR THE DICHOTOMOUS REPRESENTATION OF MACROCIRCUITS

R.M. CARPENTER and W.W. HAPP
NASA Electronics Research Center
Cambridge, Massachusetts

ABSTRACT

A compatible series of computer programs is developed for circuits with large numbers of components, typically 30-100. The programs utilize the dichotomy entailed in the flowgraph associated with any active network. Algorithms are established (i) to dichotomize the network into voltage and current generators, (ii) to examine separately each set and the associated system of controls interrelating both sets, (iii) to establish a flowgraph in terms of a unique dichotomous representation for a given equivalent circuit. While the algorithms were developed for linear newtorks, the dichotomous procedures provide an approach which is not so restricted.

1. PROBLEM FORMULATION

1.1 Need for Dichotomous Techniques

The evaluation of network functions by matrix operations such as mesh or nodal analysis has inherent limitations associated with inversion or partial inversion of matrices. Typical limitations are

(a) Necessity for substantial memory
(b) Inaccuracies due to round-off error
(c) Large running time

The complexity of a matrix-based evaluation routine increases with the number of trees inherent in the network topology, which in turn appears to increase factorially with the number of network elements. As a consequence networks with 20 to 50 components are now considered beyond the capabilities of small computers such as the IBM 1620. To meet this need several alternative avenues are being explored:

(a) Increases in computer memory from 10^7 bits to 10^{10} bits
(b) Decreases in computer operation times from 10^{-5} to 10^{-10} seconds
(c) Utilization of relaxation techniques for matrix operations
(d) Utilization of dichotomous procedures in place of matrix procedures.

Dichotomy is derived from Greek dicha (in two) and temnein (to cut) and refers to separating the system by appropriate cuts into two or more subsystems, followed by an analysis of each subsystem and of the relationships between them. Networks appear to be ideally suited for dichotomy into separate subsystems, which appear more amenable to computer-oriented methods of analysis:

(a) A subsystem relating through variables, Kirchhoff's current law
(b) A subsystem relating across variables, Kirchhoff's voltage law
(c) Interrelations between the above subsystems, Ohm's law.

The dichotomous representation of lumped parameter systems is based on functional dependence between two variables associated with each parameter, and forms the foundations of flowgraph and similar network descriptions (Ref. 1-4). This dichotomy is not related to dicoptics employed by Kron and Happ (Ref. 5).

In Kron's approach dicoptics or tearing of systems into subsystems is essentially a search for judiciously selected subsystems to minimize inter-connections and maximize intraconnections. The purpose of dicoptics is to reduce significantly the number of variables to be evaluated, while in the dichotomy entailed in flowgraphs the number of variables is doubled.

The dichotomy here proposed therefore increases the size of the associated matrix by analyzing two properties of each parameter, and aims at a reduction in the complexity of algorithms, not at a reduction in the size of the matrix associated with the system. Therefore, it is essential to distinguish between the concept of subsystem based on dichotomy defined above as separately analyzing two or more properties of each parameter of the system and the concept of subsystem resulting from dicoptics or tearing the system into parts analyzed by properties identical to those used before tearing.

1.2 Equivalent Circuit Description

A network with active and passive elements which may be non-linear can be represented graphically by an equivalent circuit as shown in Fig. 1. The equivalent circuit describes:

- A set of vertices; each vertex is defined as an electrically distinct point.
- A set of elements; each element is defined by the direction from A, the origin vertex to B, the target vertex.
- The network topology, which is defined by associating with each element a set of vertices A and B. Similarly other network properties are defined by associating with each element other properties described by the symbols C, D, E, F, G, H.
- With each element E is associated a control function C, a control element D, and a generator function G. In controlled sources $D \neq E$ and in passive elements $D = E$.
- For independent sources D is absent; this case does not occur in computer-oriented calculations and is of academic interest only.
- The dichotomy of a network is accomplished by requiring that the properties D and E be split into two mutual categories, denoted by the binary symbols 1 and 0. The dichotomy in control function D is described by property C, and the dichotomy in generator element E is described by property G.

The equivalent circuit, Fig. 2, contains the information needed to code the circuit in Fig. 1.

1.3 Review of Dichotomous Techniques

The dichotomy is based on constraints imposed by laws of physics and implemented by assigning each function a binary code 0 or 1.

- Across variables or voltages are coded by "0".
- Through variables or currents are coded by "1".
- No two voltage generators may be in parallel.
- No two current generators may be in series.
- The topology of voltage generators forms a tree or Z structure.
- The topology of current generators forms a link or Y structure.

Guidelines, algorithms and resulting programs for selection of an equivalent circuit from the schematic circuit diagram are excluded from the scope of this investigation. It is here assumed that sufficient information to construct the equivalent circuit in Fig. 2 is given. The objective of this investigation is to construct a dichotomous presentation of the network. A flowgraph is defined here as a network description in which an element is identified by:

(a) Two variables, current and voltage.

(b) A functional relation specifying the direction of functional dependence between the variables.

(c) A symbol or numerical relationship denoting this functional relationship.

The flowgraph is, therefore, a dichotomous network description, its specific form - matrix, directed graph or code - is, of course, immaterial and irrelevant.

Since the usage of the terms "block-diagram" and "flowgraph" as synonyms has lead to much confusion, it appears appropriate to define block-diagram by the following properties:

(a) An oriented weighted graph with two distinct types of nodes: Contributive or summing point; Distributive or sampling point.

(b) A functional relation between nodes specifying the direction of functional dependence between the variables.

(c) A description of the functional relationship between variables.

The flowgraph has, therefore, distinct features which distinguish it from the block-diagram in several aspects:

(a) A flowgraph is a unique description of a network while a block-diagram description entails arbitrary and subjective decisions.

(b) The functional dependence describing any network element expresses always a distributive node as a function of a contributive node, never otherwise.

(c) The functional relations governing interconnections between elements are dependent only on the network topology and are restricted to the three values -1, 0, +1. The functional dependence expresses always a contributive node in terms of distributive nodes, never otherwise.

From a draftsman's vantage point a flowgraph as defined here may, of of course, be drawn as a block-diagram without violating the above definitions of either flowgraph or block-diagram.

1.4 The Closed System Formulation

The H tag of the computer input relates the quantities to be determined by the equivalent circuit and thus defines the problem uniquely, which, in turn, is essential for the construction of a unique dichotomous solution.

For a desired circuit response, such as a transfer function, current gain or output impedance, all entries except one will be coded "0" for a real network element, with an unknown "virtual" element coded a "1". To convert the open network to a closed network by inserting a fictitious element the system is viewed as a "black box" with input and output terminals and an unknown element connecting the two terminals. The element to be determined is usually coded E = 1. The artifice of setting E(1) = 0 + j1 is often used for computation of the unknown response.

For a closed system a constraint always exists which relates the network parameters: H = 0. If the unknown is E(1) = j, then H can be expanded as

$$H(j) = H(\bar{j}) + jH(j')$$

where $H(\bar{j}) = \mathrm{Re}(H)$ and $H(j') = \mathrm{Im}(H)$

In practice H is computed as a complex number, j is redefined as an unknown and $H(j) = 0$ is solved for the unknown $j = -H(\bar{j})/H(j')$.

1.5 Illustrative Example of Network Code

The voltage regulator, Fig. 1, is coded in Table 1. The controlled source, E = 3, specifies the direction of positive current flow from node 3 to node 1. A frequency independent current source is controlled by the current through element E = 6, and thus is coded by C = 1, D = 6, F = 0, G = 1. The transmittance E = 8 with value E(8) = j1 is the unknown.

The problem solution is presented in coded form in Table 2. Thus algorithms are to be developed to produce a computer print-out from the coded network. The computer readout, Table 2, is equivalent to the flowgraph in Fig. 4.

2. ALGORITHMS FOR FLOWGRAPH CONSTRUCTION

2.1 Flowgraph Construction

The dichotomous presentation consists of three basic flowgraph regions: V, W for voltage and current, these are strictly algebraic relations; the third region T contains differential equations or subroutines, which are referred to as transmittances. To derive these relationships from the problem statement by the algorithms proceed as follows:

(1) The V matrix contains voltage relationships; the tree branches or G = 1 elements are determined in terms of the link branches or G = 0 elements.

(2) The W matrix contains current relationships of the link branches or G = 0 elements considering the tree branches or G = 1 elements as knowns.

(3) The T matrix contains the transmittances relating the W and V matrices; the known and unknown variables are determined in code G in the problem statement.

(4) The assignment of known and unknown variables is opposite in the V and W subsystems from that in the T subsystem.

The V, W, and T matrices then form a closed flowgraph.

2.2 Coding Convention for Flowgraph

The V and W matrices denote interconnections between elements or Kirchhoff constraints and the T matrix describes the intraconnection or transmittance within each element. The flowgraph is interpreted in terms of signal flow as follows:

(a) The T, V and W matrices denote signal flow from column to row entry.

(b) Signal flow in interconnections occurs from distributive to contributive nodes.

(c) Signal flow in intraconnections occurs from contributive to distributive nodes.

(d) Entries are finite; infinity is not permissible.

(e) Entries "+" or "-" imply +1 or -1; blanks imply zero.

Table 2 and Fig. 4 illustrate equivalent descriptions of a flowgraph.

2.3 Properties of V Matrix

The number of columns with entries equals the number of G = 1 elements in the equivalent circuit. Each G = 1 element is associated with one or more G = 0 elements to form a tie set, or closed sequence of adjoining elements. The algorithm must, therefore, examine the vertex origins and terminations of successive elements specified in inputs A and B, to determine which G = 1 elements must be selected to form a tie set with given G = 0 elements. From the network topology and the directions specified by A and B in the problem statement, the entries of $\underline{V}$ shown in Table 2 result.

2.4 Construction of V Matrix

Necessary data for construction are

(a) Entries A, B and G in the problem statement, Table 1.
(b) The definition of a tie associated with each G = 0 element.
(c) A definition for the sign associated with each G = 1 element in a given tie.

Establish a set of known voltages V consisting of elements E(i) for which G = 0 as:

$$V = (\ i \mid G(i) = 0\ )$$

and a conjugate set

$$\overline{V} = (\ i \mid G(i) = 1\ )$$

The set $\overline{V}$ contains only elements which are current generators, while the set V contains only voltage generators. The algorithm expresses the unknown voltages in terms of the known voltages.

$$\overline{V} = f(V)$$

There will be an entry +1 or -1 in the V(j,k) position of the $\underline{V}$ matrix if $j \in \overline{V}$ and $k \in V$. The entry will be ±1 if E(j) occurs in a tie conaining E(k). The entry will be -1 if E(j) and E(k) are in equal directions in the tie and +1 if in the opposite direction.

Select a tie of N elements with one element in $\overline{V}(j)$ and (N-1) elements in V(k) for each element in $\overline{V}(j)$. To illustrate, in Fig. 1 for the tie E(4), E(5), E(6), clearly:

$$\overline{V}(j) = \overline{V}(5)$$
$$\underline{V}(k) = (00010100)$$
$$\overline{V}(5) = V(4) + V(6)$$

This may be written as a code:

$$\overline{V}(5) = (000{+}0{+}00)$$

This result can usually be read off immediately from the equivalent circuit and is then entered as column 5 of the $\underline{V}$ matrix, or can be expressed as above as an algorithm in terms of inputs A, B and G.

2.5 Properties of W Matrix

The number of columns with entries equals the number of G = 0 elements in the equivalent circuit. Necessary data for construction are

(a) Entries A, B and G in the problem statement, Table 1.
(b) The definition of a cut-set as a single unknown G = 0 element which can be expressed in terms of one or more G = 1 elements on the basis of flow conservation through an imaginary closed surface.
(c) An assignment of sign based on the direction of flow.

Analagous to the $\underline{V}$ matrix two complementary sets exist

Unknown currents: $\overline{W} = (i \mid G(i) = 0)$
Known currents: $W = (i \mid G(i) = 1)$

The set W will contain only current generators while the set $\overline{W}$ contains only voltage generators. A set of binary coded vectors is required to express the unknown currents in terms of only the knowns.

$$\overline{W} = f(W)$$

To illustrate, consider column W6 of Table 2 which yields $W(k) = (01101011)$ and $\overline{W}(j) = (10010100)$ and the entries in the $\underline{W}$ matrix $\underline{W}(j,k) = (W(k), \overline{W}(j))$ as listed.

2.6 Sign Determinator in W(j,k)

Let $A(\overline{W})$ be the set of vertex origins of the G = 0 elements, and $B(\overline{W})$ be the set of vertex terminations of these elements. In coded form from Table 1 $A(\overline{W}) = (20030300)$ and $B(\overline{W}) = (10040100)$ where a 0 indicates a G = 0 element. To reduce the length of the code rewrite $A(\overline{W}) = N(0110)$ and $B(\overline{W}) = N(1001)$ where N(0110) is understood to be a set containing only vertices 2 and 3 of the four vertices in the equivalent circuit.

To determine the sign in W(6) separate the vertices of the equivalent circuit, Fig. 3A, into two complexes. Define complexes A(C) and B(C) as sets of vertices connected by $\overline{W}$ elements, such that E(6) is the only $\overline{W}$ element connecting the two complexes. From the problem statement Table 1 A(6) = N(0010) and B(6) = N(1000). Algorithms are formulated to identify all G = 0 elements which are connected to A(6). These are A(C) = N(0011). Similarly B(C) = N(1100).

To find all G = 1 elements connecting A(C) and B(C), connect vertices 3 or 4 in A(C) to vertices 1 or 2 in B(C). The elements E(2), E(3), E(5), E(7) and E(8) meet this requirement and are listed in Table 4 with data on A and B from Table 1. It is then determined if A(i) is in A(C) for each E(i) and B(i) is in B(C) as shown in Table 4. The result in code for E(6) is W(6) = (0+-0-0--) as entered in the W(6) position in the $\underline{W}$ matrix of Table 2, which can be interpreted as $I(6) = I(2) - I(3) - I(5) - I(7) - I(8)$.

The signs are then interpreted as the positive direction of current into A(C), the reference direction for the unknown current I(6) being specified from A(C) to B(C) or vertex A(6) = 3 to vertex B(6) = 1.

Similarly for W(4) from Table 1:

$$A(4) = N(0010) \quad \text{and} \quad B(4) = N(0001)$$

Thus

$$A(C) = N(1110) \quad \text{and} \quad B(C) = N(0001)$$

To find all G = 1 elements connecting vertices 1, 2 or 3 in A(C) to vertex 4 in B(C), note that E(5) and E(7) meet this requirement; hence W(4) = (0000+0+0).

2.7 Description of T Matrix

The $\underline{T}$ matrix describes the transmittances between the dichotomous variables, current and voltage, as shown in Table 2.

If $D \neq E$, then E is an independent source. If D = E, then E is a passive element. Four special cases occur:

(a) C = 1, G = 0 - the transmittance is an impedance Z or a current controlled voltage generator.

(b) C = 0, G = 1 - the transmittance is an admittance Y or a voltage controlled current generator.

(c) C = 0, G = 0 - the transmittance is a dimensionless quantity relating the control and generator, or a voltage controlled voltage generator.

(d) C = 1, G = 1 - the transmittance is a dimensionless quantity relating the control and generator, or a current controlled current generator.

The case D = 0 denotes that E is an independent source and is not considered here.

3. APPLICATIONS

3.1 Closure of Open Graphs

An open system, Fig. 8 , contains strictly dependent and strictly independent variables.

The flowgraph relates the variables X and Y by Y = GX, where G is the equivalent transmittance of the system. Closing the system by a "dummy" transmittance T, the "independent" variable is made G = 1/T, a function of the dependent variable X = TY. Assuming that any closed system is governed by a constraint H = 0, which is dependent on the network topology and is referred to as the topology equation, H can be expanded in terms of any parameter

$$H(T) = H(\overline{T}) + TH(T')$$

where $H(\overline{T})$ is the part of H devoid of T and H(T') is the part of H which contains T with T factored out. We then solve this equation to obtain

$$T = -H(\overline{T})/H(T') = 1/G$$

The variables X and Y are judiciously chosen; for example, to obtain current gain I(out)/I(in), choose X = I(out) and Y = I(in), so that I(in) = G I(out); thus 1/G represents the current gain.

Since present techniques of flowgraph evaluation are based on closed systems, it is necessary to specify the problem in terms of an unknown parameter T = X/Y flagged by the H tag in the problem formulation. For example, from Table 1, H(8) = 1, while H(i) = 0 for $i \neq 8$. The unknown tagged parameter is included in the flowgraph. No special algorithms are needed for processing it, provided the unknown is always treated as a transmittance associated with an element, not as a transmittance associated with a constraint due to interconnections.

3.2 Further Example of Computer Run

A second example is presented to illustrate

(a) The problem statement, including tagging of the unknown parameter.

(b) Construction of the flowgraph, including the closing parameter.

(c) Numerical values in proper form for future processing.

Fig. 5 shows the equivalent circuit of a transistorized band-pass amplifier (Ref. 6) coded for 1/G = V(out)/I(in), with Fig. 6 the coded equivalent circuit. Table 5 gives the problem statement and Table 6 the computer print-out from which the flowgraph in Fig. 7 is constructed.

Copies of this computer program have been qualified and are available at a nominal charge through project COSMIC, University of Georgia (Ref. 7).

3.3 Unique Features of Dichotomous Approach

The dichotomous approach here presented entails several original and unique features of primary interest to the systems analyst:

- A systematic procedure of problem formulation in terms of a closed system. This implies that the unknown to be evaluated forms an integral part in setting up the problem.
- An approach to circuit evaluation and design based on algorithms. This implies that the engineer formulate his problem to be understandable to a computer programmer who is not expected to acquire an understanding of the engineering aspects of the problem.
- A splitting-up of a larger system of equations into several "simpler" systems which can be analyzed with greater facility. The functional relationships describing the overall system are separated into two sets of algebraic relationships (Kirchhoff's voltage and current laws), and a set of first order differential equations relating them, if the system is linear. In non-linear systems, the linear operator is replaced by an appropriate subroutine.

REFERENCES

1. W. W. Happ and D. E. Moody, "Topological Techniques for Sensitivity Analysis", IEEE Transactions PG-ANE-11, No. 4, pp. 228-254, December, 1964.
2. W. W. Happ and J. B. Compton, "Flowgraph Models of Integrated Circuits and Film-Type Systems", IEEE Transactions on Aerospace, AS-2, pp. 259-271, 1964.
3. E. R. Robbins, D. E. Moody and W. W. Happ, "Modelling Piece-Wise Linear Systems by Topological Techniques", IEEE Transactions on Aerospace, pp. 252-264, June 1966.
4. W. W. Happ, "Flowgraphs as a Teaching Aid", IEEE Transactions on Education, pp. 69-80, July 1966.
5. H. H. Happ, "Special Cases of Orthogonal Networks", IEEE Transactions on Power Apparatus and Systems, Vol PAS-85, No. 8, August, 1966.
6. D. A. Calahan, "Computer Design of Linear Frequency Selective Networks", Proceedings of the IEEE, Vol. 53, No. 11, November 1965.
7. "A Listing of Computer Programs Available from NASA Project COSMIC", Computer Center, University of Georgia, Athens, Georgia, August 1966

Table 1. Problem Statement for Voltage Regulator (Fig. 1)

A	B	C	D	E	F	G	H	Numerical
2	1	0	0	1	0	0	0	.500 E+2
2	3	0	2	2	0	1	0	.400 E-1
3	1	1	6	3	0	1	0	.100 E+3
3	4	0	0	4	0	0	0	.250 E+2
4	1	0	5	5	0	1	0	.100 E-1
3	1	1	6	6	0	0	0	.100 E+6
4	1	0	7	7	0	1	0	.200 E-1
3	1	1	8	8	0	1	1	.100 E+1

Table 2. Problem Solution - Flowgraph Construction

V	1	2	3	4	5	6	7	8	W	1	2	3	4	5	6	7	8	T	1	2	3	4	5	6	7	8
1		+							1									1	E							
2									2	-					+			2		Y						
3									3						-			3								
4					+		+		4									4				E				
5									5			I	+		-			5			I		Y			
6		-	-		+		+	+	6									6						Z		
7									7				+		-			7							Y	
8									8						-			8								Z

Table 3. Criteria for Dichotomous Assignment of C and G

G or C	Variable	Example	G	No Elements
0	Across	Voltage	0	In parallel
1	Through	Current	1	In series

	Function	Node			
			Y	Admittance	C = 0, G = 1
C	Control	Contributive	Z	Impedance	C = 1, G = 0
			W	Amplification	C = 1, G = 1
G	Generator	Distributive	V	Amplification	C = 0, G = 0

Table 4. Choice of Entries for W Matrix

E(i)	A(i)	B(i)	Is A in A(C)?	Is B in B(C)?	±
2	2	3	No	No	+
3	3	1	Yes	Yes	-
5	4	1	Yes	Yes	-
7	4	1	Yes	Yes	-
8	3	1	Yes	Yes	-

Table 5. Problem Statement for Band-Pass Amplifier (Fig. 5)

A	B	C	D	E	F	G	H	Numerical
2	1	1	1	1	0	0	0	.500 E+2
2	1	0	2	2	1	1	0	.400 E-5
2	1	0	3	3	2	1	0	.250 E-4
2	3	0	4	4	0	1	0	.500 E+2
3	1	1	5	5	0	0	0	.500 E+8
3	1	0	6	6	1	1	0	.200 E-8
3	4	0	7	7	1	1	0	.500 E-11
4	1	0	5	8	0	1	0	.500 E+2
4	1	0	9	9	0	1	0	.400 E+4
4	1	0	10	10	1	1	0	.100 E-8
4	1	1	11	11	2	0	0	.100 E-2
1	2	0	11	12	0	1	1	.100 E+1

Table 6. Computer Output for Flowgraph Construction

```
 V 1 2 3 4 5 6 7 8 9 10 11 12  W 1 2 3 4 5 6 7 8 9 10 11 12
 1   + + +                  +  1
 2                             2 -
 3                             3 -
 4                             4 -       +
 5       -   + +               5
 6                             6         -
 7                             7         -             +
 8                             8                       -
 9                             9                       -
10                            10                       -
11             - - + +        11
12                            12 +

               T 1 2 3 4 5 6 7 8 9 10 11 12
               1 Z
               2   Y
               3     Y
               4       Y
               5         Z     Y
               6           Y
               7             Y
               8
               9                 Y
              10                    Y
              11                       Z  Y
              12
```

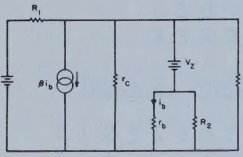

R1
V1
βib
rc
Vz
ib
rb
R2
RL

FIG. 1 Equivalent circuit for voltage regulator

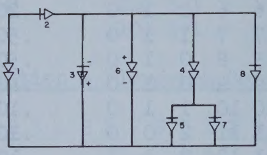


FIG. 2 Coded equivalent circuit, exhibiting dichotomy

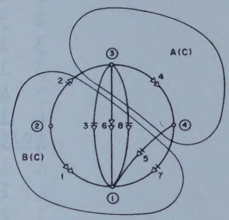

A(C)
B(C)
③
②
④
①
1 2 3 4 5 6 7 8

FIG. 3A Separation of equivalent circuit into cut-set for the evaluation of I (6)

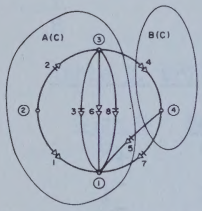

A(C)
B(C)
③
②
④
①
1 2 3 4 5 6 7 8

FIG. 3B Separation of equivalent circuit into cut-set for the evaluation of I (4)

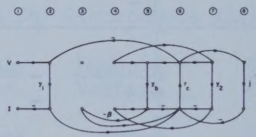

① ② ③ ④ ⑤ ⑥ ⑦ ⑧
V
I
y1
yb
rc
y2
j
-β

FIG. 4 Flowgraph constructed from V, W, and T matrices

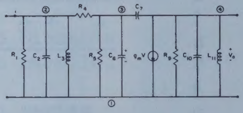

② ③ ④ ①
R1 C2 L3 R4 R5 C6 C7 gmV R9 C10 L11 V0

FIG. 5 Band-pass amplifier equivalent circuit

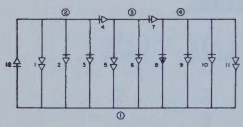

② ③ ④ ①
12 1 2 3 4 5 6 7 8 9 10 11

FIG. 6 Coded equivalent circuit showing dichotomy

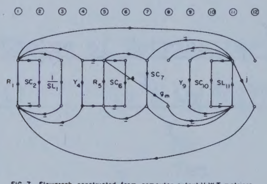

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫
R1 SC2 1/SL1 Y4 R5 SC6 SC7 gm Y9 SC10 SL11 j

FIG. 7 Flowgraph constructed from computer output V, W, T matrices

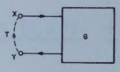


FIG 8. Closed System Representation, Y=GX, X=TY

MAXIMALLY FLAT APPROXIMATION TECHNIQUES

H. J. ORCHARD
Lenkurt Electric Co., Inc.
San Carlos, California

G. C. TEMES
Ampex Corporation
Redwood City, California

ABSTRACT

The operations permissible in power-series equivalences are analyzed. Based on the theorems thus established, practical design procedures are developed for the approximation of prescribed network characteristics in the neighbourhood of an arbitrary frequency. The desired characteristics may represent the modulus of a transfer function, the loss, the phase or delay, or a combination of these. The approximation criterion is a maximally flat error-function.

The procedure leads to simultaneous linear equations in the coefficients of the desired transfer function. If $\underline{k}$ of the derivatives of the specified network characteristics are arbitrary, then an $\underline{nth}$ order transfer function can be used to match n + k derivatives. The suitable values of the $\underline{k}$ unspecified derivatives can be found from the vanishing of $\underline{k}$ determinants, each of order n + 1. These define the $\underline{k}$ derivatives as the eigenvalues of a system of n + k linear equations in the $\underline{n}$ unknown coefficients. Iterative methods can be used to find the eigenvalues.

The application of these techniques to loss, phase, as well as simultaneous loss and phase approximation, is demonstrated. As specific examples, the maximally flat approximation of a linear phase, in a lowpass and bandpass interval, is described. The equivalence between the new solutions thus found and those previously described is proven.

A GENERAL CLASS OF TUNABLE ACTIVE R-C FILTERS

J. V. Leeds, Jr.
Rice University
Houston, Texas

ABSTRACT

In the design of low frequency filters, the solution is often an active R-C filter. If a requirement exists that the filter be tunable, then until recently only a few specific R-C circuits, e.g. (Twin-Tee, Wein-Bridge) were known to be tunable by varying the resistors. Recent work has shown that a very general class of R-C circuits may be tuned by varying the resistors. The sensitivity of these filters to parameter tolerances is examined here and shown to be independent of the tuning problem.

An appendix also discusses a circuit in which a co-tree capacitor current is a state variable of the network in addition to the normal tree capacitor voltages. This circuit arose during the derivation of the tunable filter conditions and is of interest in its own right.

INTRODUCTION

The tunable R-C active/passive network is defined as a network whose frequency response can be varied uniformly by varying uniformly the resistors in that network. The conditions for this to be fulfilled are discussed elsewhere in detail (1). Briefly, the conditions are that the network contain no independent current sources, that the tree capacitor voltages be the only state variables, and that all the resistors be varied uniformly. Naturally if resistors are varied, one is interested in determining what effect this has on the sensitivities of the network responses. This problem is examined here. The result is that the normalized sensitivities of the voltage response values depend only on the nominal value of the resistors and are time or frequency scaled as is the original response values. The normalized current sensitivities do not have this property but are varied in proportion to the resistor variation. Finally, a circuit is presented in which a tree capacitor voltage and a co-tree capacitor current are the state variables.

ANALYSIS

As has been known for some time, the response of an R-C network is time or frequency scaled if all the capacitors are modified by multiplying them by some constant. Let the normalized sensitivity be defined as

$$S_N = a \frac{\partial Y}{\partial a}$$

where Y is some response variable (voltage or current).

a is some network element (R or C).

Since the response of the network is only time or frequency scaled, one would suspect that S_N would be only time or frequency scaled. Such is the case and will be proved shortly.

Since the tunable active/passive networks are tuned by varying the resistors, it is natural to ask if S_N in this case is only frequency or time scaled but not changed otherwise. The answer is yes for some variables but no for others. In the "no" case, the change is easily related to the resistor variation. Now a proposition will be proved which gives the relations between the normalized sensitivities and the frequency range.

In order to understand the proof, it is necessary to discuss the method of calculating network sensitivity previously developed (2). The method consists of constructing an auxiliary network and calculating the response of this network. This network is a duplicate of the original network except that all sources are set equal to zero. Then, a single source is used to drive the auxiliary network. This is a current source and it is placed across one element. The auxiliary network response variables are the normalized sensitivities of the corresponding variables in the original network with respect to the element which has the current source in parallel with it. The current source is equal to the current flowing in the original network in that element.

The analysis of the auxiliary network for the tunable R-C filter is now developed. The results of the analysis of the tunable active filter are also summarized. Since the auxiliary network is driven by current sources, we need only the analysis of that type of source. The auxiliary network variables are designated by $*$.

Consider a general connected R-C network with no controlled sources. Reference (1) shows that if controlled sources are allowed, the following analysis in general is still valid provided that V_{CT} are the only state variables of network. Let the following vector quantities be defined:

V_S = Independent Voltage Sources

I_S = Independent Current Sources

V_{CT} = Voltages across Capacitors in the Tree

V_{CC} = Voltages across Capacitors in the Co-tree

V_{RT} = Voltages across Resistors in the Tree

V_{GC} = Voltages across Resistors in the Co-tree

I_{CT} = Currents through Capacitors in the Tree

I_{CC} = Currents through Capacitors in the Co-tree
I_{RT} = Currents through Resistors in the Tree
I_{GC} = Currents through Resistors in the Co-tree
I_{VS} = Current through Voltage Sources
V_{IS} = Voltages across Current Sources
D_{CT} = Capacitors in Tree
D_{CC} = Capacitors in Co-tree
D_{RT} = Resistors in Tree
D_{GC} = Resistors in Co-tree
B_{ij} = Submatrix of Fundamental Circuit Matrix
S_{ij} = Submatrix of Fundamental Cutset Matrix

Let a tree be selected containing all the voltage sources, as many of the capacitors as possible, and as many resistors as necessary. The tree is selected in that order: voltage source, capacitor, resistor. Then the following equations result:

$$\begin{pmatrix} I_{VS} \\ I_{RT} \\ I_{CT} \end{pmatrix} + \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & 0 & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix} \begin{pmatrix} I_{GC} \\ I_{CC} \\ I_{S} \end{pmatrix} = 0 \tag{1}$$

$$\begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & 0 & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix} \begin{pmatrix} V_{S} \\ V_{RT} \\ V_{CT} \end{pmatrix} + \begin{pmatrix} V_{GC} \\ V_{CC} \\ V_{IS} \end{pmatrix} = 0 \tag{2}$$

$$D_{CT} \dot{V}_{CT} = I_{CT} \tag{3}$$

$$D_{CC} \dot{V}_{CC} = I_{CC} \tag{4}$$

$$D_{RT} I_{RT} = V_{RT} \tag{5}$$

$$D_{GC} V_{GC} = I_{GC} \tag{6}$$

These are the equations of a general R-C network. For the case of no voltage source, one has $B_{11}, B_{21}, B_{31}, S_{11}, S_{12}, S_{13}$ identically equal to zero. This is the case in the auxiliary network. We now proceed to derive the equations of the auxiliary network. One first eliminates $\overset{*}{I}_{RT}$ and $\overset{*}{V}_{GC}$ from (5) and (6) using (1) and (2):

$$\overset{*}{V}_{RT} = -D_{RT}(S_{21} \overset{*}{I}_{GC} + S_{23} \overset{*}{I}_{S}) \tag{7}$$

$$\overset{*}{I}_{GC} = -D_{GC}(B_{12} \overset{*}{V}_{RT} + B_{13} \overset{*}{V}_{CT}) \tag{8}$$

Thus one has:

$$\overset{*}{I}_{GC} = -D_{GC}(-B_{12} D_{RT} \overset{*}{S}_{21} I_{GC} + B_{13} \overset{*}{V}_{CT} - B_{12} D_{RT} S_{23} \overset{*}{I}_{S}) \tag{9}$$

or

$$(U - D_{GC}\ B_{12}\ D_{RT}\ S_{21}) I^*_{GC} = -D_{GC}(B_{13}\ V^*_{CT} - B_{12}\ D_{RT}\ S_{23}\ I^*_S) \qquad (10)$$

Now is each resistor is assumed to have an initial value and the actual value of each resistor is changed proportionally by this value, one has

$$D_{RT} = K\ D_{RTo} \qquad (11)$$

$$D_{GC} = \frac{1}{K}\ D_{GCo} \qquad (12)$$

where K is a scalar variable.

Thus Equation (10) becomes

$$\begin{aligned} &(U - D_{GCo}\ B_{12}\ D_{RTo}\ S_{21}) I^*_{GC} \\ &= -\frac{1}{K}\ D_{GCo}(B_{13}\ V^*_{CT} - K\ B_{12}\ D_{RTo}\ S_{23}\ I^*_S) \end{aligned} \qquad (13)$$

or

$$Q^{-1}\ I^*_{GC} = -\frac{1}{K}\ D_{GCo}(B_{13}\ V^*_{CT} - K\ B_{12}\ D_{RTo}\ S_{23}\ I^*_S) \qquad (14)$$

where $Q = (U - D_{GCo}\ B_{12}\ D_{RTo}\ S_{21})^{-1}$.

Note Q is not a function of K.

Now using (2) in (4), V_{CC} is eliminated:

$$-D_{CC}\ B_{23}\ \dot{V}^*_{CT} = I^*_{CC} \qquad (15)$$

Next using (9), (14), (15), (1) in (3), only I^*_S, and V^*_{CT} remain:

$$\begin{aligned} D_{CT}\ \dot{V}^*_{CT} &= -(S_{31}\ I^*_{GC} + S_{32}\ I^*_{CC} + S_{33}\ I^*_S) \\ &= -S_{31}\ \frac{Q}{K}\ D_{GCo}\ (B_{13}\ V^*_{CT} - K\ B_{12}\ D_{RTo}\ S_{23}\ I^*_S) \\ &\quad + S_{32}\ D_{CC}\ (B_{23}\ \dot{V}^*_{CT}) - S_{33}\ I^*_S \end{aligned} \qquad (16)$$

or

$$\begin{aligned} K(D_{CT} - S_{32}\ D_{CC}\ B_{23})\ \dot{V}^*_{CT} &= -S_{31}\ Q\ D_{GCo}\ B_{13}\ V^*_{CT} \\ &\quad - K\ S_{31}\ Q\ D_{GCo}\ B_{12}\ D_{RTo}\ S_{23}\ I^*_S \\ &\quad - S_{33}\ K\ I^*_S\ . \end{aligned} \qquad (17)$$

Equation (17) is the state equation of the system.

Now consider the back solution. First the currents:

$$I^*_{GC} = \frac{1}{K}\ \{-Q\ D_{GCo}\ B_{13}\ V^*_{CT}\} \qquad (18)$$

$$I^*_{RT} = \frac{1}{K}\ \{-S_{21}\ Q\ D_{GCo}\ B_{13}\ V^*_{CT}\} \qquad (19)$$

$$I_{CT} = \frac{1}{K}\ \{-S_{31}\ Q\ D_{GCo}\ B_{13}\ V^*_{CT} + K\ S_{32}\ D_{CC}\ B_{23}\ \dot{V}^*_{CT}\} \qquad (20)$$

$$I^*_{CC} = \frac{1}{K}\ \{K\ D_{CC}\ B_{23}\ \dot{V}^*_{CT}\} \qquad (21)$$

Now consider the voltage equations. First:

$$V^*_{RT} = D_{RTo}\ S_{21}\ Q\ D_{GCo}\ B_{13}\ V^*_{CT} \qquad (22)$$

$$V^*_{CC} = -B_{23}\ V^*_{CT} \qquad (23)$$

$$V^*_{GC} = -B_{12}\, V^*_{RT} - B_{13}\, V^*_{CT} \tag{24}$$

Equations 17 through 24 describe the response of the auxiliary network. In this Case, I_S contains one and only one non-zero entry. That entry is the current through the element that the sensitivities being determined with respect to.

We now summarize the network equations for the tunable R-C network (1):

$$K(D_{CT} - S_{32}\, D_{CC}\, B_{23})\, \dot{V}_{CT} = -S_{21}\, Q\, D_{GCo}(B_{11}\, V_S + B_{13}\, V_{CT}) + K\, S_{32}\, D_{CC}\, B_{21}\, \dot{V}_S \tag{25}$$

$$V_{RT} = D_{RTo}\, S_{21}\, Q\, D_{GCo}\, (B_{11}\, V_S + B_{13}\, V_{CT}) \tag{26}$$

$$V_{CC} = -B_{21}\, V_S - B_{23}\, V_{CT} \tag{27}$$

$$V_{GC} = -B_{11}\, V_S - B_{12}\, V_{RT} - B_{13}\, V_{CT} \tag{28}$$

$$I_{GC} = \frac{1}{K}\, \{-Q\, D_{GCo}\, (B_{11}\, V_S + B_{13}\, V_{CT})\} \tag{29}$$

$$I_{RT} = \frac{1}{K}\, \{-S_{21}\, Q\, D_{GCo}\, (B_{11}\, V_S + B_{13}\, V_{CT})\} \tag{30}$$

$$I_{CT} = \frac{1}{K}\, \{-S_{31}\, Q\, D_{GCo}\, (B_{11}\, V_S + B_{13}\, V_{CT}) + K\, S_{32}\, D_{CC}(B_{21}\, \dot{V}_S + B_{23}\, \dot{V}_{CT})\} \tag{31}$$

$$I_{CC} = \frac{1}{K}\, \{K\, D_{CC}(B_{21}\, \dot{V}_S + B_{23}\, \dot{V}_{CT})\} \tag{32}$$

Thus, as noted previously, everywhere a derivative appears and only where a derivative appears in the voltage equations (25-28) does a K appear. Hence, varying K varies the frequency response of the network by the factor $1/K$.

Now examine Equations 17 through 24. Everywhere that the term I_S appears, the term K multiplies it. But everywhere in the current equations of the original network (29-33), the factor $1/K$ appears. Thus suppose one has for $K = 1$, calculated the response of the original network and determined the sensitivities of the voltages in the auxiliary network. Then, suppose $K \neq 1$, and one determines the response of the original network. The currents in that network will be time (frequency) scaled and reduced by the value $1/K$. But everywhere I_S appears in the auxiliary network, it is multiplied by K. Thus the excitation to the auxiliary network is the same as the case $K = 1$ except that it is time (frequency) scaled. Thus the voltages of the auxiliary network will be the same for $K = 1$ or $K \neq 1$ except that for $K \neq 1$, they will be time scaled. However, all the current sensitivity equations (18, 19 and 20) are divided by K. Hence, the normalized current sensitivities for $K = 1$

are divided by K for $K \neq 1$. As shown previously (1), these results apply to active networks in which V_{CT} are the only state variables of the network since the same basic equations, 1-6, describe this type of active network.

Now consider varying the capacitors to tune a network. The capacitors only appear in the equations, 25, 31 and 32. Hence we need only to note that any proportional factor can be absorbed in the time scale. Hence, varying the capacitors does not change any value of current or voltage. Thus, the current sensitivities of the auxiliary network are only time scaled.

Now fractional sensitivity is defined as

$$S_f = \frac{a}{Y} \frac{\partial Y}{\partial a}$$

(Note in the time domain, this may be undefined at the points of time where $Y = 0$. This is the reason for using the normalized sensitivity on the past sections.) If one examines the equations given above, one will see that all the currents in the original network are divided by $1/K$, and similarly all the current sensitivities in the auxiliary network are also divided by $1/K$. Hence, the S_f for currents in the auxiliary network do not contain K, and S_f is only time-frequency scaled.

Thus, we have the following proposition:

Given an arbitrary, connected R-C network containing dependent current and voltage sources, independent voltage sources; but no independent current sources, with V_{CT} as the only state variables, then multiplying all the resistors by a factor K will result in:

(a) The network response (current and voltage) will be scaled in frequency by $1/K$ or scaled in time by K.

(b) All voltages will remain the same except from time or frequency scaling.

(c) All the currents will be multiplied by $1/K$.

(d) The sensitivities, both fractional and normalized, of the response voltages of the network will be the same except for frequency or time scaling.

(e) The normalized sensitivities of the response currents of the network will be multiplied by $1/K$ and time or frequency scaled. The fractional sensitivities of the response currents will be time or frequency scaled only.

Also, given an arbitrary R-C network containing dependent and independent current and voltage sources, multiplying all the capacitors by a factor M will result in

(a) The network response, both current and voltage, will be time or frequency scaled.

(b) The sensitivities, both normalized and fractional, will be similarly time or frequency scaled and otherwise unchanged.

A similar result will hold for the R-L case with independent current sources as drivers.

DISCUSSION

Consider now the problem of design of a tunable active filter. One naturally needs to set tolerances on the components. Suppose, the sensitivities varied with the tuning of the network other than being frequency-time scaled. In this case, one would have to examine the tolerance problem over the entire tuning range. However, this is not the case. A single evaluation of the sensitivities of the network for one setting of the resistors is sufficient to determine the tolerances over the entire tuning range. This results in quite a simplification of the problem of determining parameter tolerances.

ACKNOWLEDGMENT

This research was supported in part by NSF Grants GK-262 and GU-1153.

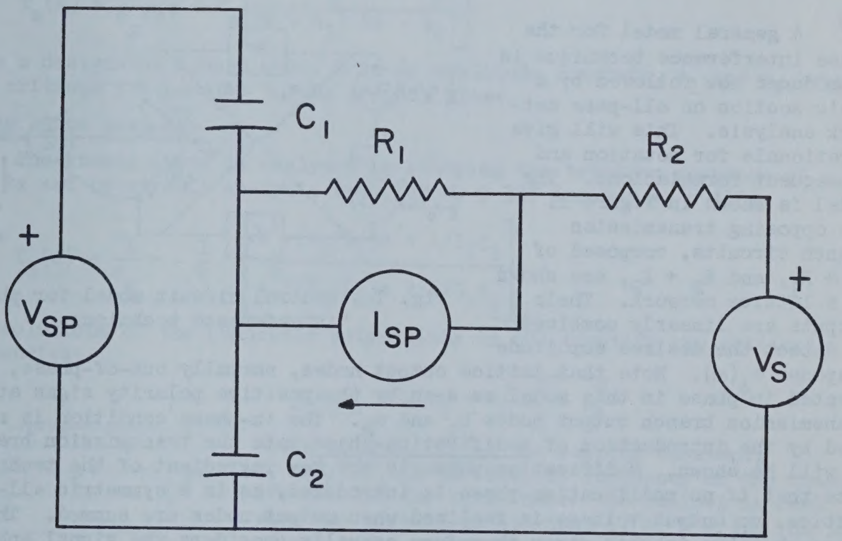
BIBLIOGRAPHY

(1) J. V. Leeds, "Design of Tunable R-C Active Filters and Variable Bandwidth Band Pass-Band Elimination Filters", July 1966, EE-66-4, Department of Electrical Engineering, Rice University.

(2) J. V. Leeds, "Transient and Steady State Sensitivity Function", IEEE Transactions on Circuit Theory, CT-13-3, September 1966.

APPENDIX A

As stated in the main body, one restriction on the results is that V_{CT} be the only state variables. Figure A1 shows a network which has as state variables a tree capacitor voltage and a co-tree capacitor current. A straight forward analysis of that current will show that V_{C2} and I_{C1} or I_{C2} and V_{C1} are state variables of that network.



$$I_{SP} = k\, I_{C_2}$$

$$V_{SP} = m\, V_{R_1}$$

A NOVEL APPROACH IN THE DESIGN OF AUDIO PHASE INTERFERENCE FILTERS USING HIGHLY DAMPED CIRCUITS

LAWRENCE V. KRIGER
Air Force Cambridge Research Laboratories, Bedford, Mass.

ABSTRACT

Phase interference techniques have been successfully employed in the audioband using two opposing, highly damped, passive networks to synthesize various filter characteristics. For this, a nonsymmetric lattice network is treated as a split channel transmission system, each branch of which is a highly damped, all-pass circuit. A prescribed phase-delay is synthesized in each branch after which the two are linearly combined to phase-detect a desired frequency response. Selective filter characteristics with minimized transient duration may be obtained using few components. A novel synthesis technique is discussed for realizing - 18 db/octave high-pass and low-pass filter sections. This response may also include a discrete rejection frequency at no additional circuit complexity.

I. GENERAL MODEL

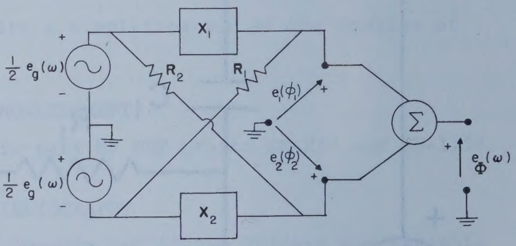


Fig. 1 - General circuit model for phase interference technique.

A general model for the phase interference technique is introduced now followed by a basic section on all-pass network analysis. This will give a rationale for notation and subsequent formulations. The model is shown in Figure 1. The opposing transmission branch circuits, composed of $X_1 + R_1$, and $R_2 + X_2$, are shown as a lattice network. Their outputs are linearly combined to detect the desired amplitude response $e_\Phi(\omega)$. Note that lattice output nodes, normally out-of-phase, are treated in-phase in this model as seen by the positive polarity signs at transmission branch output nodes e_1 and e_2. The in-phase condition is realized by the introduction of modification-phase into the transmission branches as will be shown. Modification-phase is the key ingredient of the technique. Note that if no modification-phase is introduced, as in a symmetric all-pass lattice, no output voltage is realized when output nodes are summed. The approach taken in this paper therefore actually considers the signal enhancement properties gained by use of modification-phase rather than the complementary view of signal attenuation by phase-interference, although either description could have been used. Both approaches are phase interference techniques.

A transmission branch is composed of a pure reactance and a pure resistance. When the reactance arm is connected to positive input reference, it defines a mode 1 transmission (see $X_1 + R_1$). When the resistive arm is connected there, it is in mode 2 (see $X_2 + R_2$). The general model may thus be functionally described as a split channel network of two phase-modified all-pass transmission branches that are linearly combined to phase detect a preselected frequency band.

II. ANALYSIS OF BASIC ALL-PASS NETWORKS

The transfer response of some all-pass transmission branches will now be derived and parameters analyzed with respect to their phase modifying characteristics. To review, an all-pass function (Ref 1-4) is a sub class of non-minimum-phase functions that are characterized by zeros in the right half of the s plane. A non-minimum-phase function is considered all-pass if only zeros are in the right hand plane and only poles in the left which are mirror images of the zeros about the $j\omega$ axis. The magnitude of an all-pass-function response is constant, independent of applied frequency. The phase response of all-pass networks derived herein exhibit a monotonically decreasing slope as a function of increasing frequency and regardless of mode. Two such phase responses are realized for a given filter section. Their relationship is the basis for the phase interference technique.

The voltage transfer response of a second order all-pass function is pertinent and therefore now basically discussed. Its overdamped RLC circuit case is detailed since it is used in the examples to minimize transient duration. The system function may be written as:

$$T_m(s) = \frac{e_m(s)}{e_g(s)} = K_m\left[\frac{(s + s_1)\ (s + s_2)}{(s - s_1)\ (s - s_2)}\right] \tag{1}$$

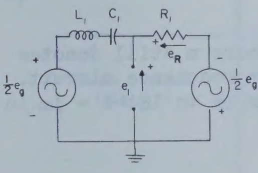
where m designates branch mode, K is an amplitude constant, s_1 and s_2 represent critical frequencies in the complex plane.

Steady State Response

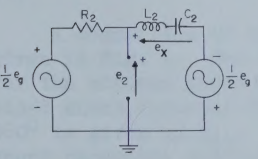
The steady state is analyzed by studying the transfer response of Fig. 2a and by circuit analysis (see Ref 2, p 337 and Ref 4, p 315).

$$T_1(s) = \frac{e_1}{e_g} = -\frac{1}{2}\left[\frac{s^2-(R_1/L_1)s + 1/L_1C_1}{s^2+(R_1/L_1)s + 1/L_1C_1}\right] \tag{2}$$

Root solutions of the quadratic polynomials in (2) identify the critical frequencies:



a. Mode 1



b. Mode 2

Fig. 2 - RLC branch circuit modes.

$$\pm s_{1,2} = \mp \alpha_1 \pm \sqrt{\alpha_1^2 - \omega_1^2} \tag{3}$$

where $\alpha_1 = R_1/2L_1$ is the damping constant, $\omega_1 = 1/L_1C_1$ is the natural frequency of the reactive arm and the term under the radical is real when the circuit is overdamped, $(\alpha_1 > \omega_1)$. The loci of pole-zero paths for various degrees of damping (ω_1 fixed) are similar to the minimum-phase case except for an added folded over path in the right half plane for the zeros. In this paper we will be mainly concerned with the case $(\alpha_m \geq \omega_n)$ wherein the pole-zero paths lie on the real axis. For equal component values in each circuit of Fig. 2 it is obvious that:

$$T_2(s) = -\,T_1(s);\quad \text{(equal components)} \tag{4}$$

It should now also be obvious, for steady state consideration, that output amplitude has a constant value of 1/2 and phase, as a function of frequency, is described by appropriately summing the angles of the pole and zero phasors of (1) with the argument of the constant multiplier.

Transient Response

The impulse response may be found by rewriting (2) in a form appropriate for conversion (see Ref 3, p 440):

$$T_1(s) = -\frac{1}{2}\left[\frac{s^2 - 2\alpha_1 s + \omega_1^2}{s^2 + 2\alpha_1 s + \omega_1^2}\right] = -\frac{1}{2}\left[1 - \frac{4\alpha_1 s}{(s - s_1)(s - s_2)}\right]. \qquad (5)$$

Inverse Laplace transformation gives the impulse response of the series RLC all-pass network:

$$f_0(t) = \mathcal{L}^{-1}\, T_1(s) = -\frac{1}{2}\left[u_0(t) - \frac{4\alpha_1}{s_2 - s_1}\left(s_2 e^{s_2 t} - s_1 e^{s_1 t}\right)u_{-1}(t)\right] \qquad (6)$$

where the impulse in the response is due to the all-pass nature of the network. The step function response is found by integrating (6):

$$f_1(t) = -\left[\frac{1}{2} + \frac{\alpha_1}{\sqrt{\alpha_1^2 - \omega_1^2}}\left(e^{s_2 t} - e^{s_1 t}\right)\right] u_{-1}(t) \qquad (7)$$

In an underdamped circuit, the ringing duration after an applied step depends on how little damping exists in the circuit. For an overdamped step response, the duration of the exponential transient depends on how large the damping. For practical purposes in phase interference techniques, it has been found that minimum transient duration effects occur near critical damping.

Phase Response

Although the phase response of an overdamped all-pass RLC circuit was analyzed from the phasor relations described by (1), its phase angle is best calculated from the magnitudes of its real and imaginary arms:

$$\phi_{mn}(\omega) = -2 \tan^{-1} \frac{X_{mn}}{R_m} \qquad (8)$$

At this point, subscript notation is more elaborate, where m still denotes mode number and n is an integer relating the particular reactance circuit. A direct method of assessing X_{mn} and R_m in compact form is to let $s = j\omega$ in (2) and then separate real and imaginary parts.

$$T_{10}(\eta) = -\frac{1}{2}\left[\frac{1 + j\,\dfrac{1}{q_{10}(\eta - 1/\eta)}}{1 - j\,\dfrac{1}{q_{10}(\eta - 1/\eta)}}\right] \qquad (9)$$

The simplifying notation used in (9) to express T_{mn} is:

$$\eta = \omega/\omega_{mn}\ ;\ q_{mn} = \omega_{mn} L_{mn}/R_m$$

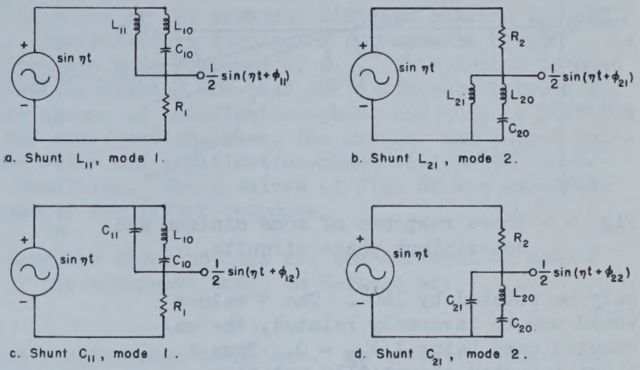
where η is a normalizing frequency parameter and q is a classic constant, but written in lower case, since there are no classic Q or resonant effects in overdamped circuits. The phase response, (derived in a previous paper [Ref 5] by the author) is obtained from (9) as:

$$\phi_{10} = -2 \tan^{-1} q_{10}(\eta - 1/\eta) \tag{10}$$

From previous considerations (4), the alternate mode response of (10) is:

$$\phi_{20} = -2 \tan^{-1} q_{20}(\eta - 1/\eta) \pm \pi \tag{11}$$

The modification of a branch phase response in this paper will only involve the addition of a shunt inductor or capacitor across the series LC arm. Modified circuits are noted in Fig. 3. Their modified phase responses are derived in a manner similar to that for (10) and (11) and are given in Table 1 for mode 1 circuits.



+
−
sin ηt
L_{11}
L_{10}
C_{10}
R_1
$\frac{1}{2}\sin(\eta t+\phi_{11})$
R_2
L_{21}
L_{20}
C_{20}
$\frac{1}{2}\sin(\eta t+\phi_{21})$
C_{11}
$\frac{1}{2}\sin(\eta t+\phi_{12})$
C_{21}
$\frac{1}{2}\sin(\eta t+\phi_{22})$

a. Shunt L_{11}, mode 1. b. Shunt L_{21}, mode 2.

c. Shunt C_{11}, mode 1. d. Shunt C_{21}, mode 2.

Fig. 3 - Circuits for extending phase range.

The mode 2 responses are obtained by simply adding ± π radians. In Table 1, the new symbol γ_{mn} represents the ratio of series to shunt inductance or capacitance. The addition of a phase-modifying shunt component across the series LC arm introduces a pole which may be designed to appear outside the band of interest and its transient effect (at a branch output) is usually small.

Mode 1 Circuit	Phase Response	Description of Symbols
Series RLC (Fig. 2a)	$\varphi_{10} = -2 \tan^{-1} q_{10}(\eta - 1/\eta)$	$q_{10} = \frac{\omega_{10}L_{10}}{R_1}$; $\omega_{10} = \frac{1}{L_{10}C_{10}}$
Shunt L Added (Fig. 3a)	$\varphi_{11} = -2 \tan^{-1} \frac{q_{11}(\eta - 1/\eta)}{1 + (1 - 1/\eta^2)\gamma_{11}}$	$q_{11} = \frac{\omega_{11}L_{10}}{R_1}$; $\omega_{11} = \frac{1}{L_{10}C_{10}}$ $\gamma_{11} = L_{10}/L_{11}$
Shunt C Added (Fig. 3c)	$\varphi_{12} = -2 \tan^{-1} \frac{q_{12}(\eta - 1/\eta)}{1 - \frac{(\eta^2 - 1)}{\gamma_{12}}}$	$q_{12} = \frac{\omega_{12}L_{10}}{R_1}$; $\omega_{12} = \frac{1}{L_{10}C_{10}}$ $\gamma_{12} = C_{10}/C_{11}$; $\eta = \omega\sqrt{L_{10}C_{10}}$

TABLE 1. Transmission Branch Phase Formulas for Mode 1.

A comparison of minimum and non-minimum phase network responses is given in Fig. 4. For a series RLC circuit, the phase asymptotes are ± 90° for the minimum phase case and ± 180° for the non-minimum. The effect of an added shunt inductor for increasing the low frequency asymptote from 180° to 360° is given by ϕ_{11}. Similar but inverse considerations apply for the shunt capacitor case of ϕ_{12} where the high frequency asymptote is extended to -360°. By additional reactance circuit complexity, phase range may be further extended. The effect of γ value in modifying phase response, for the case of a shunt inductor, is shown in Fig. 5. The unshunted case is given by $\gamma_{11} = 0$. For the case of a shunt capacitor, Fig. 5 curves need

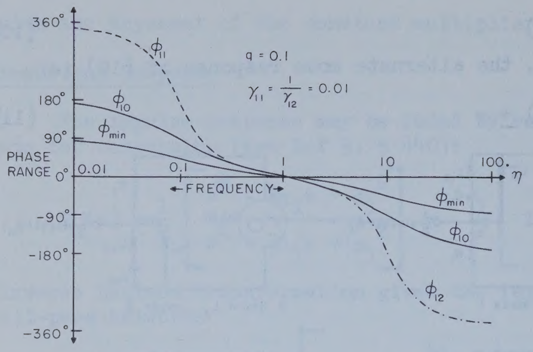


Fig. 4 - Phase response of some minimum and non-minimum phase circuits.

only be rotated by 180°. The γ values would now be inversely related, the unshunted case being $1/\gamma_{12} = 0$. Thus a shunt inductor essentially modifies phase response below $\eta = 1$, and a shunt capacitance above.

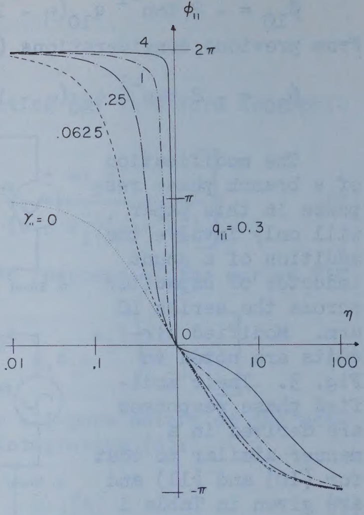
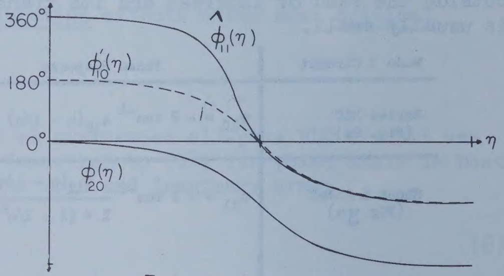


Fig. 5 - Effect of γ_{11} in extending phase range at low frequencies.

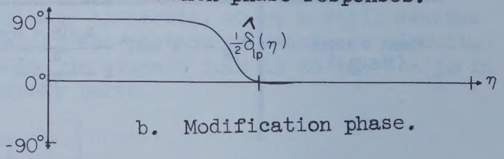
III. APPROXIMATING A LOW-PASS AND HIGH-PASS FILTER RESPONSE, SIMPLE ESTIMATE

Comprehension of the synthesis technique is simplified somewhat by first analyzing a particular simple approach. The derived parameters pertinent to this special case will be symbolized with an overhat. The overhat will be omitted if the extracted parameter is the same in the general case to be derived. This simple estimate will be a special version of the general case.

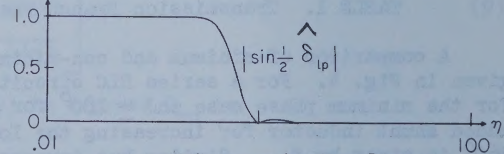
The following three notions must be recognized and understood in order to follow the synthesis procedure. First, the description of a transmission branch phase in terms of its alternate mode. Second, the derivation of an expression for modification-phase employing this notion. Third, the use of the expression for modification-phase to express the amplitude response of a filter network. These three notions are simply portrayed in Fig. 6 to obtain a low-pass filter response.



a. Branch phase responses.



b. Modification phase.



c. Filter output voltage.

Fig. 6 - Phase techniques for synthesizing low pass filter response, simple estimate.

The branch phase responses that apply are ϕ_{11} and ϕ_{20}. Recognize that ϕ_{20} is for a series RLC circuit in mode 2 and ϕ_{11} is for a mode 1 shunt-L case (see Fig. 3). For frequencies above $\eta = 1$, the two curves are seen in Fig. 6a to be approximately 180° out-of-phase. At very low frequencies they become in-phase (or 360° out-of-phase). If we describe modification-phase to be a modulus of the in-phase condition between ϕ_{11} and ϕ_{20}, it can be seen that this notion is captured by the vertical distance between ϕ_{11} and ϕ_{10}'. The ϕ_{10}' curve was constructed by displacing ϕ_{20} upwards by 180°. As such it represents the alternate mode response of ϕ_{20}. This notion is used in the formula for modification-phase. One-half the difference between ϕ_{11} and ϕ_{10}' is defined as the amount of modification-phase and this is pictured in Fig. 6b. The larger the modification-phase, the greater the output voltage response. The largest value for modification-phase will be 90° corresponding to the no loss condition. The δ values of Fig. 6b are employed, as shown in Fig. 6c, to obtain the filter response.

The formulas which describe these notions are now derived. A mode 2 phase response in terms of its alternate mode, is defined as:

$$\phi_{2n} = \phi_{1n}' \pm \pi \tag{12}$$

where the prime designates displacement technique. Modification-phase may be derived by a trigonometric treatment of both branch phase responses to an arbitrary input signal $\sin\eta t$. The filter output voltage would then be:

$$e_{\Phi}(\eta) = 1/2 \sin(\eta t + \phi_{1n}) + 1/2 \sin(\eta t + \phi_{2n}) \tag{13}$$

Inserting the alternate expression for ϕ_{2n} of (12) into (13) we get by trigonometric identity:

$$e_{\Phi}(\eta) = \sin\frac{\phi_{1n} - \phi_{1n}'}{2} \quad \cos\left[\eta t + \frac{\phi_{1n}}{2} + \frac{\phi_{1n}'}{2}\right] \tag{14}$$

Examination of (14) identifies the sine term as amplitude coefficient and the cosine term as frequency. Concerning ourselves with magnitude only:

$$|\, e_{\Phi}(\eta)\,| = |\, \sin\frac{\phi_{1n} - \phi_{1n}'}{2}\,| = |\, \sin\frac{\delta(\eta)}{2}\,| \tag{15}$$

from which the definition for modification-phase is obtained:

$$\frac{\delta(\eta)}{2} \triangleq \frac{\phi_{1n} - \phi_{1n}'}{2} \tag{16}$$

The response of (15) is constrained to describe a filter characteristic in the following example.

A simple estimate will now be made for deriving a low pass filter response. The technique will try to constrain (15) to a response slope which falls off at - 18 db/octave. A polynomial which closely expresses this fall off rate is $1/\eta^3$. The problem now is to approximate this polynomial by constraining (15) to equal $1/\eta^3$ in a filter's rejection band and also to equal unity in the pass band. Unity is obtained by maintaining the low frequency asymptotic phase responses of ϕ_{1n} and ϕ_{1n}' near 180° and 360°. The solution to the problem requires that:

$$\sin\frac{\delta(\eta)}{2}\Big|_{\eta\gg 1} \approx \tan\frac{\delta(\eta)}{2} \approx \frac{1}{\eta^3}\, ; \quad \sin\frac{\delta(\eta)}{2}\Big|_{\eta < 1} \approx 1 \tag{17}$$

Since the argument of the sine function becomes very small in the filter's rejection band, the tangent form may be substituted. This allows the derived tangent expressions for branch phase to be incorporated into a formula for modification-phase. Thus for this special example, a filter response may be approximated by:

$$\tan \hat{\delta}_{1p}(\eta)/2 = \tan(\hat{\phi}_{11} - \phi_{10}')/2 \tag{18}$$

By trigonometric identify, (18) may be expanded to:

$$\tan\frac{\hat{\delta}_{1p}(\eta)}{2} = \frac{\tan\hat{\phi}_{11}(\eta)/2 - \tan\phi_{10}'(\eta)/2}{1 + (\tan\hat{\phi}_{11}(\eta)/2)(\tan\phi_{10}'(\eta)/2)} \qquad (19)$$

Revising the arctangent expressions for branch phase for reinsertion into (19) gives a rather cumbersome expression. This may be alleviated somewhat by describing $\hat{q}_{11}$ from the constraint that the numerator approach zero as frequency becomes infinite:

$$\hat{q}_{11} = q_{20}(1 + \gamma_{11}) \qquad (20)$$

Substituting the expression for $\hat{q}_{11}$ of (20) back into the cumbersome expression results in one less unknown. This is then manipulated into a polynomial form from which further analytic constraints may be applied. To facilitate analytic considerations at this point, the polynomial coefficients are alphabetically symbolized:

$$\tan\frac{\hat{\delta}_{1p}(\eta)}{2} = \frac{\hat{C}(\eta^2 + \hat{D})}{\eta(\eta^4 + \hat{B}\eta^2 + \hat{A})} \qquad (21)$$

where

$$\hat{A} = 1 - \frac{\gamma_{11}}{q_{20}^{\,2}(1 + \gamma_{11})}\ ;\ \hat{B} = \frac{1}{q_{20}^{\,2}}\ ;\ \hat{C} = \frac{\gamma_{11}}{q_{20}(1 + \gamma_{11})}\ ;\ \hat{D} = -1$$

An enhanced third order difference is obtained between numerator and denominator by setting $\hat{A} = 0$ in (21);

$$\tan\frac{\hat{\delta}_{1p}(\eta)}{2}\bigg|_{\hat{A}=0} = \frac{\hat{C}(\eta^2 + \hat{D})}{\eta^3(\eta^2 + \hat{B})} \ (22); \quad \gamma_{11} = \frac{q_{20}^{\,2}}{1 - q_{20}^{\,2}} \qquad (23)$$

The constraint of A = 0 was satisfied by (23). Finally, it would be desirable in (22) to approach $1/\eta^3$ by setting $\hat{B} = \hat{D}$, but this has no practical solution. (This is realized in the general case.) This therefore is as far as we can go in approximating (19) to a $1/\eta^3$ polynomial and the results are now analyzed. To do this, various values of q_{20} are assumed and the other parameters ($\hat{\gamma}_{11}$ and $\hat{q}_{11}$) computed. The value of q_{20} must be limited to less than one since for $q_{20} \geq 1$, γ_{11} of (25) takes on negative values. The resulting responses are given in Fig. 7. The results indicate a large error at η=1. The frequency of infinite attenuation at $\eta = 1$ cannot be avoided as long as ϕ_{11} and ϕ_{10}' both have zero phase at $\eta = 1$ as was seen in Fig. 6a. The null could be avoided if ϕ_{11} were sufficiently displaced to the right (or ϕ_{10}' to the left). This simple technique thus inherently inserts a notch in a filter envelope characteristic. A monotonic -18 db/octave low pass characteristic is only obtained, however, for frequencies above $\eta = 2$, where amplitude loss is already severe, or at very low frequencies.

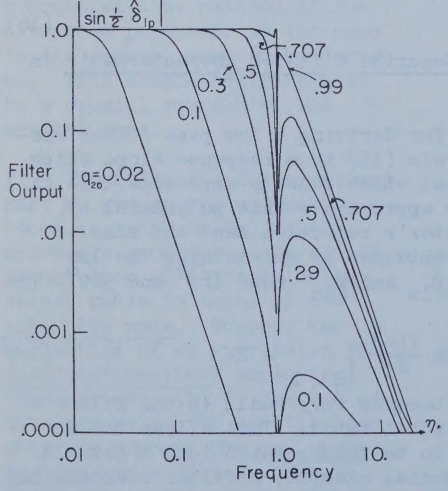


Fig. 7 - Low-pass filter envelope response for various $\hat{q}$-$\hat{\gamma}$ estimates.

The inverse high-pass case gives equal response curves if the frequency axis is inverted. The difference in derivation is that the polynomial to be approximated for the high-pass case of -18 db/octave is η^3 (instead of $1/\eta^3$). The constraint that would now be used for deriving constants and variables is that (19) approach zero, as η approaches zero frequency.

IV. APPROXIMATING LOW-PASS AND HIGH-PASS FILTER SECTIONS, GENERAL CASE

The general case technique for synthesizing -18 db/octave filter sections essentially follows the previous section. Only one innovation is introduced, a horizontal phase displacement constant λ. As was previously indicated, displacing one of the transmission branch phase responses along the log frequency axis makes it possible to eliminate the response dip at $\eta = 1$. Now, for the low-pass case, ϕ_{11} will be shifted to the right by a factor λ_{11}. The phase displaced curve is derived by rewriting the response formula for ϕ_{11} in terms of a displaced normalized frequency:

$$\eta' = \eta/\lambda_{mn} \tag{24}$$

The expression for modification-phase now becomes:

$$\frac{\delta_{1p}(\eta)}{2} = \frac{\phi_{11}(\eta') - \phi_{10}'(\eta)}{2} \tag{25}$$

where the prime over η in the first numerator term of (25) indicates horizontal displacement along the frequency axis, and the second term's prime indicates π radian vertical displacement along the phase axis. Again using the -18 db/octave low-pass requirement as an example, we are required to constrain

$$\tan \left.\frac{\delta_{1p}(\eta)}{2}\right|_{\eta \gg 1} = \frac{\tan \phi_{11}(\eta')/2 - \tan \phi_{10}'(\eta)/2}{1 + \tan \phi_{11}(\eta')/2 \tan \phi_{10}'(\eta)/2} \approx \frac{1}{\eta^3} \tag{26}$$

Again expanding (26), rewriting η' expressions in terms of η and solving for q_{11} as before:

$$q_{11} = q_{20}\lambda_{11}(1 + \gamma_{11}) \tag{27}$$

LOW-PASS CASE	HIGH-PASS CASE
$\tan \frac{\delta_{lp}}{2} = - \frac{C^2(\eta^2 + D)}{\eta(\eta^4 + B\eta^2 + A)}$	$\tan \frac{\delta_{hp}}{2} = - \frac{H\eta^3(G\eta^2 + 1)}{E\eta^4 + F\eta^2 + 1}$
$A = \lambda_{11}^2 \left[1 - \frac{\gamma_{11}}{q_{20}^2(1 + \gamma_{11})}\right]$	$E = \frac{1}{\lambda_{12}^2} \left[1 - \frac{1}{q_{20}^2(1 + \gamma_{12})}\right]$
$B = \frac{1}{q_{20}^2} - 1 - \lambda_{11}^2$	$F = \frac{1}{q_{20}^2} - \frac{(1 + \lambda_{12}^2)}{\lambda_{12}^2}$
$C = \frac{1}{q_{20}} \left[1 - \frac{\lambda_{11}^2}{1 + \gamma_{11}}\right]$	$H = \frac{1}{q_{20}} \left[1 - \frac{\gamma_{12}}{\lambda_{12}^2 (1 + \gamma_{12})}\right]$
$D = -\left[\frac{\lambda^2\gamma_{11}}{1 + \gamma_{11} - \lambda_{11}^2}\right]$	$G = -\left[\frac{1}{\gamma_{12}\lambda_{12}^2 + \lambda_{12}^2 - \gamma_{12}}\right]$

Table 2 Coefficients for transcendental expressions of modification-phase, general case.

Substituting for q_{11} and rearranging terms gives the desired polynomial:

$$\tan \frac{\delta_{1p}(\eta)}{2} \Bigg|_{\eta \gg 1} = - \frac{C(\eta^2 + D)}{\eta(\eta^4 + B\eta^2 + A)} \tag{28}$$

The letter coefficients of (28) are described in Table 2. Now, after a similar constraining procedure as in the previous case, the low-pass parameter constants are derived and results are given in Table 3 in terms of q_{20}. The results of a -18 db/octave high-pass derivation are also given for comparison purposes.

LOW-PASS CASE	HIGH-PASS CASE
$Y_{11} = \dfrac{q_{20}^2}{1 - q_{20}^2}$	$Y_{12} = \dfrac{1 - q_{20}^2}{q_{20}^2}$
$\lambda_{11} = \left[\dfrac{1}{2q_{20}^2}\left(1 - \sqrt{1 - 4q_{20}^2}\right)\right]^{1/2}$	$\lambda_{12} = \left[\dfrac{1}{2}\left(1 + \sqrt{1 - 4q_{20}^2}\right)\right]^{1/2}$
$q_{11} = \dfrac{1}{(1 - q_{20}^2)}\left[\dfrac{1}{2}\left(1 - \sqrt{1 - 4q_{20}^2}\right)\right]^{1/2}$	$q_{12} = \dfrac{1}{(1 - q_{20}^2)}\left[\dfrac{1}{2}\left(1 - \sqrt{1 - 4q_{20}^2}\right)\right]^{1/2}$

Table 3 Parameter solutions for -18 db/octave attenuation slopes, general case.

Analyzing Table 2, note the inverse frequency relationship of high and low-pass letter coefficients seen by inverting them in either tangent expression. More directly, by inverting λ and δ in the expression for any given letter constant, gives the form of its inverse case, e.g., between A and E, etc. This is confirmed in Table 3 where the parameter constants between high and low-pass cases are easily seen to be inversely related. In the special case, the constraint B = D could not be realized. It is realizable here and was satisfied by the expressions for λ given in the table. For circuit q_{20} = 0.3, the value of λ_{11} is calculated to be 1.054. This shifting value may be appreciated analyzing the family of response curves in Fig. 8. Here the special case of null at η = 1 is given by the curve labeled λ_{11} = 1 (no horizontal displacement of ϕ_{11}). Some arbitrary values of λ_{11} are plotted to show the transition to the optimum -18 db/octave slope where λ_{11} = 1.054. This response curve accurately approaches - 18 db/octave, no rejection frequency being encountered.

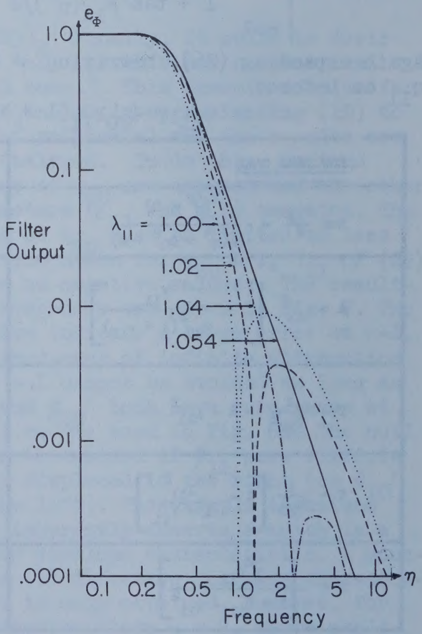
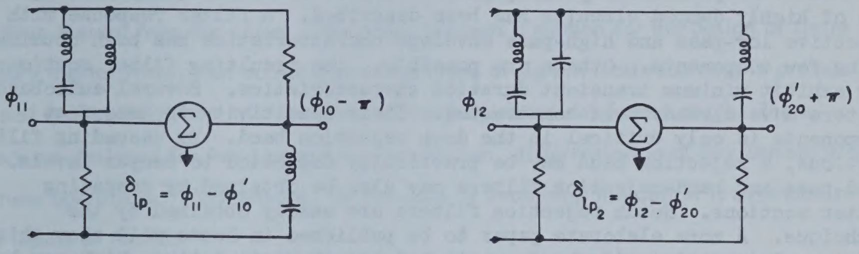


Fig. 8 - Effect of λ_{11} on envelope response.

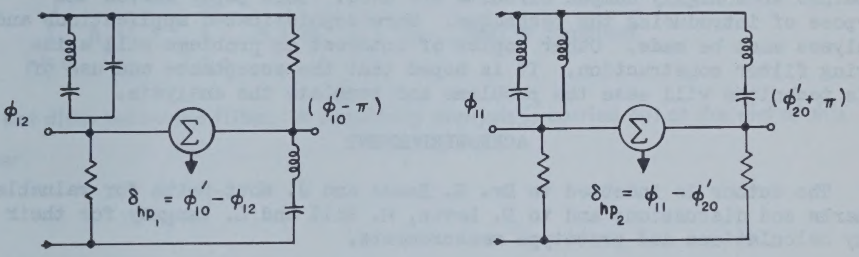
An alternate model results if one of the modes in the general model of Fig. 1 is turned around. Here, a low-pass characteristic becomes high-pass and vice versa after such a switch around. The four circuits possible from the two models are shown in Fig. 9. The modification-phase related to each circuit is pictorially described in Fig. 10. Here it may be observed that the low-pass filter circuit of Fig. 9a with modification-phase δ_{lp_1} may be simply switched to the circuit of Fig. 9d to obtain a high-pass response of modification-phase δ_{hp_2}. Their adjacent modification-phase regions are given by the two upper areas in Fig. 10. Similar considerations apply for δ_{hp_1}



ϕ_{11}
$(\phi_{10}' - \pi)$
$\delta_{lp_1} = \phi_{11} - \phi_{10}'$
ϕ_{12}
$(\phi_{20}' + \pi)$
$\delta_{lp_2} = \phi_{12} - \phi_{20}'$

a. General model lowpass filter

c. Alternate model lowpass filter



ϕ_{12}
$(\phi_{10}' - \pi)$
$\delta_{hp_1} = \phi_{10}' - \phi_{12}$
ϕ_{11}
$(\phi_{20}' + \pi)$
$\delta_{hp_2} = \phi_{11} - \phi_{20}'$

b. General model high pass filter

d. Alternate model high pass filter

Fig. 9 - Notation for general and alternate circuit models.

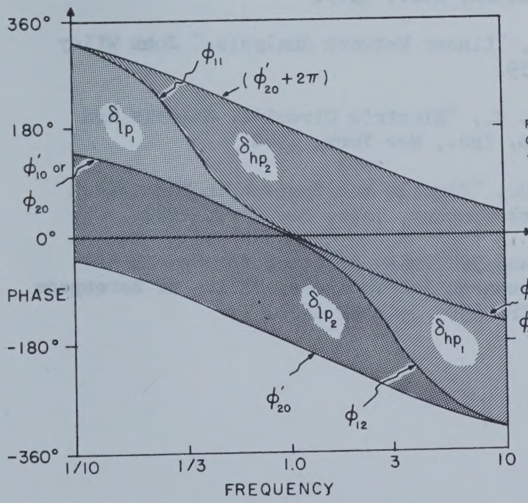


Fig. 10 - Modification-phase areas.

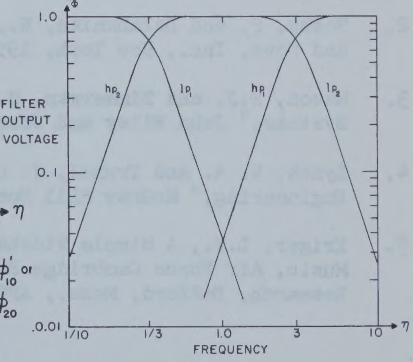


Fig. 11 - Low-pass and high-pass -18db/octave characteristics for general and alternate circuit models.

and δlp_2. The filter envelope responses are seen in Fig. 11. Note how switched responses will intersect at their 3 db down corner frequency. The -3db frequency is found by solving for $\delta/2 = 45^\circ$. The corner frequencies may occur at $\eta = 1/3$ or at $\eta = 3$ when $q_{20} = 0.3$. By analyzing a particular area in Fig. 10, it may be appreciated that large modification-phase values in or near the pass band may be practically achieved with very small error, while requirements deep in a rejection band are exceedingly difficult.

SUMMARY

Network synthesis employing phase interference techniques allowing the use of highly damped circuits has been described. A filter response with selective low-pass and high-pass envelope characteristics has been obtained using few components. Others are possible. The resulting filter sections may exhibit minimum transient duration characteristics. Several audioband filters have already been constructed. Their sensitivity to imperfect components is only critical in the deep rejection band. By cascading filter sections, a rejection band may be practically depressed to deeper levels. Band-pass and band-rejection filters may also be obtained by cascading filter sections. Notch rejection filters are easily obtained by the technique. A more elaborate paper to be published in-house will show this and some interesting wide-band constant-delay characteristics which may be obtained when highly damped circuits are used. This paper served the purpose of introducing the technique. More sophisticated applications and analyses must be made. Other topics of interest or problems will arise during filter construction. It is hoped that the acceptance and use of this technique will ease the problems and complete the analysis.

ACKNOWLEDGEMENT

The author is indebted to Dr. K. Haase and J. Mott-Smith for valuable remarks and discussions and to D. Levin, M. Hill and L. Tanguay for their many calculations and prototype measurements.

REFERENCES

1. Bode, H. W., "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., Princeton, N.J., 1945.

2. Seshu, S. and Balabanian, N., "Linear Network Analysis," John Wiley and Sons, Inc., New York, 1959.

3. Mason, S.J. and Zimmerman, H. J., "Electric Circuits, Signals and Systems," John Wiley and Sons, Inc., New York, 1960.

4. Lynch, W. A. and Truxal, J. G., "Signals and Systems in Electrical Engineering," McGraw Hill Book Company, Inc., New York, 1962.

5. Kriger, L.V., A Simple Wideband 90° Phase Splitter for Speech or Music, Air Force Cambridge Research Laboratories, Office of Aerospace Research, Bedford, Mass., AFCRL 930, October 1961.

ON A CLASS OF FINITE-MEMORY DIGITAL FILTERS

GEORGE SHURAYM
Texas Instruments Incorporated
Dallas, Texas

It is important in the design of process control systems that the filters used do not distort the signal at their input and at the same time extract this signal or some function thereof from the noise. This paper presents a method for the design of finite memory digital filters that do not distort the types of signals most recurring in process control systems (exponential, sinusoidal, decaying sinusoidal forms, etc.). These filters are designed for input smoothing, prediction, differentiation or any combination of these operations. Furthermore, for the case of smoothing the design is extended so that inputs of the form

$$\sum_{i=o}^{J} B_i t^i + \sum_{i=o}^{M} A_i t^{k_i} e^{-\alpha_i t} \quad (\alpha_i, \text{ real or complex})$$

are not distorted by the filter. A sensitivity analysis is carried out at the end of this paper.

INSERTION-LOSS ANALYSIS AND EXACT DESIGN OF THE EQUAL-ELEMENT CLASS OF TEM TRANSMISSION-LINE FILTERS

HARVEY J. HINDIN AND JESSE J. TAUB

Airborne Instruments Laboratory
Applied Electronics Division
Melville Long Island, New York

ABSTRACT

The insertion-loss characteristics of a class of TEM transmission-line filters have been derived for any number of resonators. The filters are formed by quarter-wave shorted-stub resonators or half-wave open-stub resonators all having the same characteristic impedance and coupled by a quarter-wavelength of transmission line. This constitutes the "equal-element" design, the advantages of which are described in the literature. They may be successfully used in many cases of interest and do not require the tapering of stub impedances that is necessary in the Butterworth or Tchebycheff designs. The results are presented in graphical form for one through eight resonators and a wide range of stub impedances, and are used in the insertion-loss analysis and synthesis of this class of filter. A discussion of the insertion loss equations and curves presented, which represent functions of the Tchebycheff polynomials of the second kind, is made to improve the overall understanding of these filters. It is shown that they possess useful low-pass, band-pass, and band-stop properties. Bandwidths, impedances, number of resonators, etc. can be rigorously determined for specific cases of interest using the material presented.

THE SYNTHESIS OF ORTHOGONAL FILTERS BY THE USE OF THE STURM-LIOUVILLE EQUATIONS

MICHEAL PARTEN and R. H. SEACAT
Texas Technological College
Lubbock, Texas

ABSTRACT

The orthogonal properties of the eigenfunctions of the proper Sturm-Liouville problem are utilized to give a series representation of any bounded piecewise-linear transfer function. The series solution to the Sturm-Liouville equation is used as a general set of eigenfunctions. The independent variable is changed to extend the region of orthogonality from zero to infinity. Rearranging the new series results in a series of exponentials which can be realized as simple RC networks with amplifiers representing the coefficients. Since the general eigenfunctions used are equivalent to most of the familiar orthogonal functions, only one particular orthogonal function need be used to determine the coefficients necessary to represent a given transfer function. The Legendre polynomials are used for this purpose. An arbitrary constant is generated in the change of variable process which is directly related to the RC time constants and indirectly related to the amplifier gains. A few examples are given to show the effect of different RC time constants and of a different number of terms in the series on the amplifier gains and on the accuracy of the series representation of the transfer function. The RC networks with amplifiers will, in many cases, be able to be used in integrated circuit form.

MORGAN-VOYCE POLYNOMIALS IN RECURRENT LADDERS

M.N.S. SWAMY, MEMBER, IEEE and B. B. BHATTACHARYYA
Nova Scotia Technical College
Halifax, Nova Scotia

ABSTRACT

Ladder networks have been previously analysed by a number of methods such as continued fractions, Tchebycheff polynomials, iteration, signal flow graphs, continuants, etc. Morgan-Voyce demonstrated in the special case of a resistive ladder network, an interesting mathematical relationship between some of the network functions and the Fibbonacci series of numbers and defined a set of polynomials by

$$b_n(w) = wB_{n-1}(w) + b_{n-1}(w) \qquad (n \geqq 1)$$

$$B_n(w) = (w + 1)\ B_{n-1}(w) + b_{n-1}(w) \qquad (n \geqq 1)$$

with

$$b_o(w) = B_o(w) = 1$$

In subsequent papers, Swamy studied in detail, the properties of these polynomials and defined the polynomial $c_n(w)$ as

$$c_n(w) = \frac{1}{2}\ (b_n + b_{n-1}) = \frac{1}{2}\ (B_n - B_{n-2})$$

In the present paper these polynomials b_n, B_n and c_n are used to make a detailed study of recurrent ladders with a general series impedance z_1 and a shunt impedance z_2. The [ABCD] parameters have been expressed in terms of these polynomials.

Since the polynomials are factorisable, the matrix parameters can be conveniently utilised to yield network response to any arbitrary excitation. Also, the resonant and anti-resonant frequencies of the network may be found in terms of the zeros of these polynomials.

For purposes of illustration, responses to a step and a square wave input have been worked out in detail in the case of an RCG ladder. The steady state response of the RCG ladder for the periodic excitation has also been obtained from the total response.

The matrix parameters of a uniform line have also been obtained by considering it to be the limiting case of a recurrent ladder. The chief merit of the analysis lies in its simplicity and compactness.

TRAJECTORY SENSITIVITY OF AN OPTIMAL CONTROL SYSTEM

HENRY D'ANGELO*
MAYNARD L. MOE* and
THOMAS C. HENDRICKS**

ABSTRACT

A method is introduced for analytically determining the optimal control for a linear system with respect to a performance functional that includes trajectory sensitivity. Specifically, the control law u is found for the system described by the state-variable differential equation

$$\dot{x} = Ax + Bu$$

so that the performance index J is minimized:

$$J = \int_0^\infty \left(x^T Q x + u^T Ru + \sum_{i=1}^{m} v_i^T S_i v_i \right) dt$$

where Q, R and the S_i are positive definite matrices and the trajectory sensitivity vector v_i is defined by

$$v_i = \frac{\partial^i x}{\partial \alpha^i}$$

where α is a system parameter. Numerical determination of the optimal-control law is achieved by the solution of a nonlinear algebraic matrix equation. Implementation of the optimal control requires generation of the v_i vector which is done in a straight-forward manner: for an n-th order system, the trajectory sensitivity vector v_i requires n integrators for its simulation.

1. The Main Results

Consider the plant

$$\dot{x} = Ax + Bu \qquad (1)$$

where

$A = A(\alpha)$, $B = B(\alpha)$ and

$$x = x(t, \alpha) = \begin{bmatrix} x_1(t, \alpha) \\ \cdot \\ \cdot \\ \cdot \\ x_n(t, \alpha) \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_r \end{bmatrix}$$

---

* Associate Professor of Electrical Engineering at the University of Denver.

** Senior Research Scientist at the Martin Company in Denver and graduate student at the University of Denver.

Assuming a nominal α_0 the Taylor series expansion can be written about this value as

$$x(t,\alpha) = x(t,\alpha_0) + \frac{\partial x(t,\alpha_0)}{\partial \alpha}(\alpha-\alpha_0) + \frac{\partial^2 x(t,\alpha_0)}{\partial \alpha^2}\frac{(\alpha-\alpha_0)^2}{2!} + \dots$$
$$= x(t,\alpha_0) + v_1(t,\alpha_0)(\alpha-\alpha_0) + v_2(t,\alpha_0)\frac{(\alpha-\alpha_0)^2}{2!} + \dots \tag{2}$$

where

$$v_i(t,\alpha) = \frac{\partial^i x(t,\alpha)}{\partial \alpha^i} \tag{3}$$

Clearly if the trajectory $x(t, \alpha)$ is to be independent of α it is necessary that

$$v_i(t,\alpha) = 0 \text{ for } i = 1, 2, 3, \dots \tag{4}$$

Generally, for any significant system parameter α, making the $v_i(t, \alpha) = 0$ will conflict with the system's intended function. A performance index which allows one to weight trajectory deviations relative to the control effort and the state transfer in the "regulator" problem is

$$J = \int_0^\infty \left(x^T Qx + u^T Ru + \sum_{i=1}^\infty v_i^T S_i v_i \right) dt = \text{minimum} \tag{5}$$

For practical purposes it is assumed that $x(t, \alpha)$ is well approximated by a finite number of terms. Thus equation (2) can be replaced by

$$x(t,\alpha) = x(t,\alpha_0) + \sum_{i=1}^m v_i(t,\alpha_0)\frac{(\alpha-\alpha_0)^i}{i!} \tag{6}$$

Thus the performance index J becomes

$$J = \int_0^\infty \left(x^T Qx + u^T Ru + \sum_{i=1}^m v_i^T S_i v_i \right) dt \tag{7}$$

Taking derivatives repeatedly with respect to α of equation (6) results in

$$v_j(t,\alpha) = 0, \quad j = m+1, m+2, \dots \tag{8}$$

A new state vector X is formed by adjoining the system state variable vector x with the trajectory sensitivity vectors $v_1, \dots v_m$. Thus

$$X = [x^T, v_1^T, \dots, v_m^T]^T \tag{9}$$

and has dimension $(m+1)n$. In terms of this new state vector X, the performance index J, given by equation (7), can be written as

$$J = \int_0^\infty (X^T \hat{Q} X + u^T Ru)\, dt \tag{10}$$

where

$$\hat{Q} = \begin{bmatrix} Q & 0 & \cdots & 0 \\ 0 & S_1 & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots & S_m \end{bmatrix} \tag{11}$$

For the present, let it be assumed that this new state vector X satifies the differential equation

$$\dot{X} = \hat{A}X + \hat{B}u \tag{12}$$

If this assumption is true, then the optimal control u minimizing the performance index J of equation (10) is known to be

$$u = -R^{-1}\hat{B}^T KX \tag{13}$$

where K satisfies the Riccati equation

$$K\hat{A} + \hat{A}^T K - K\hat{B}R^{-1}\hat{B}^T K + \hat{Q} = 0 \tag{14}$$

Substituting the optimal control u of equation (13) into the original plant equation (1) gives

$$\dot{x} = Ax - BR^{-1}\hat{B}^T KX \tag{15}$$

A more convenient form of the above equation is obtained by partitioning the K matrix:

$$K = \begin{bmatrix} K_{00} & K_{01} & K_{02} & \cdots & K_{0m} \\ K_{01}^T & K_{11} & K_{12} & \cdots & K_{1m} \\ K_{02}^T & K_{12}^T & K_{22} & \cdots & K_{2m} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ k_{0m}^T & K_{1m}^T & K_{2m}^T & \cdots & K_{mm} \end{bmatrix} \tag{16}$$

where the K_{ii} are symmetric for $i = 0, \ldots, m$. Also

$$K = [K_0, K_1, K_2, \ldots, K_m] \tag{17}$$

where

$$K_i = [K_{0i}^T, K_{1i}^T, \ldots, K_{i-1,i}^T K_{ii}, K_{i,i+1}, \ldots, K_{i,m-1}, K_{im}]^T,$$

$$i = 0, 1, 2, \ldots, m \tag{18}$$

Equation (15) can now be written

$$\dot{x} = Ax - BR^{-1}\hat{B}^T [K_0, K_1, \ldots, K_m][x^T, v_1^T, \ldots, v_m^T]^T \tag{19}$$

or

$$\dot{x} = (A - BR^{-1}\hat{B}^T K_0)x - BR^{-1}\hat{B}^T \sum_{i=1}^{m} K_i v_i \tag{20}$$

Equation (20) is written in the compact form

$$\dot{x} = A_{00}x + \sum_{i=1}^{m} A_{0i}\, v_i \tag{21}$$

where

$$A_{00} = A - BR^{-1}\hat{B}^T K_0 \tag{22}$$

and

$$A_{0i} = -BR^{-1}\hat{B}^T K_i, \; i = 1, \ldots, m \tag{23}$$

Equations (22) and (23) can be combined into the single equation

$$A_{0i} = \delta_{0i} A - BR^{-1}\hat{B}^T K_i, \; i = 0, 1, \ldots, m \tag{24}$$

where δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{cases} 1, & i=j \\ 0, & i\neq j \end{cases} \tag{25}$$

Taking the derivative of equation (21) with respect to α gives a differential equation for v_1:

$$\dot{v}_1 = A_{10}x + \sum_{i=1}^{m} A_{1i}\, v_i + A_{1,\,m+1}\, v_{m+1} \tag{26}$$

where the coefficients A_{1i}, $i = 0, 1, \ldots, m$ are discussed in the sequel. In accordance with Equation (8)

$$v_{m+1} = 0 \tag{27}$$

Therefore equation (26) reduces to

$$\dot{v}_1 = A_{10}x + \sum_{i=1}^{m} A_{1i}\, v_i \tag{28}$$

Proceeding in a similar manner results in the general form

$$\dot{v}_j = A_{j0}x + \sum_{i=1}^{m} A_{ji}\, v_i, \; j = 1, \ldots, m \tag{29}$$

Adjoining equations (1) and (29) results in the partitioned matrix equation

$$\begin{bmatrix} \dot{x} \\ \dot{v}_1 \\ \cdot \\ \cdot \\ \cdot \\ \dot{v}_m \end{bmatrix} = \begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ A_{10} & A_{11} & A_{12} & \cdots & A_{1m} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ A_{m0} & A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix} \begin{bmatrix} x \\ v_1 \\ \cdot \\ \cdot \\ \cdot \\ v_m \end{bmatrix} + \begin{bmatrix} B \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} u \tag{30}$$

or simply

$$\dot{X} = \hat{A}X + \hat{B}u \tag{31}$$

where

$$\hat{A} = \begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ A_{10} & A_{11} & A_{12} & & A_{1m} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ A_{m0} & A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix} \tag{32}$$

and

$$\hat{B} = [B^T, 0, \ldots, 0]^T \tag{33}$$

It is now possible to determine $\hat{A}$ and $\hat{B}$ in terms of known matrices and partitions of K. Substituting from equations (18) and (33) in equation (22) gives

$$A_{00} = A - BR^{-1}B^T K_{00} \tag{34}$$

Similarly, substituting from equations (18) and (33) in equation (23) gives

$$A_{0i} = -BR^{-1}B^T K_{0i}, \; i = 1, \ldots, m \tag{35}$$

Equations (34) and (35) can be combined into the single equation

$$A_{0i} = \delta_{0i} A - BR^{-1}B^T K_{0i}, \; i = 0, 1, \ldots, m \tag{36}$$

where δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{cases} 1, & i=j \\ 0, & i=j \end{cases} \tag{37}$$

To simplify the ensuing formulations the following notation is introduced:

$$A^{(k)} = \frac{\partial^k A}{\partial a^k} \tag{38}$$

$$\beta^{(k)} = \frac{\partial^k (-BR^{-1}B^T)}{\partial a^k} \tag{39}$$

Thus equations (21) and (38) can be written as

$$\dot{x} = A_{00} x + \sum_{i=1}^{m} A_{0i} v_i \tag{40}$$

where

$$A_{0i} = \delta_{0i} A^{(0)} + \beta^{(0)} K_{0i}, \; i = 0, \ldots, m \tag{41}$$

Taking the derivative of equation (40) with respect to α results in

$$\dot{v}_1 = A_{10} x + \sum_{i=1}^{m} A_{1i} v_i \tag{42}$$

where

$$\begin{aligned} A_{1i} &= \theta(i-1)\, A_{0,\, i-1} + \frac{\partial A_{0i}}{\partial \alpha} \\ &= \theta(i-1)\, [\delta_{0,\, i-1}\, A^{(0)} + \beta^{(0)} K_{0,\, i-1}] \\ &\quad + [\delta_{0i}\, A^{(1)} + \beta^{(1)} K_{0i}], \quad i = 0, \ldots, m \end{aligned} \tag{43}$$

where $\theta(t)$ is the unit-step function defined by

$$\theta(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{44}$$

Taking the derivative of equation (42) with respect to α results in

$$\dot{v}_2 = A_{20} x + \sum_{i=1}^{m} A_{2i} v_i \tag{45}$$

where

$$\begin{aligned} A_{2i} &= \theta(i-2)\, [\delta_{0,\, i-2}\, A^{(0)} + \beta^{(0)} K_{0,\, i-2}] \\ &\quad + 2\,\theta(i-1)\, [\delta_{0,\, i-1}\, A^{(1)} + \beta^{(1)} K_{0,\, i-1}] \\ &\quad + [\delta_{0i}\, A^{(2)} + \beta^{(2)} K_{0i}], \quad i = 0, \ldots, m \end{aligned} \tag{46}$$

Proceeding in this manner results in

$$\dot{v}_j = A_{j0} x + \sum_{i=1}^{m} A_{ji} v_i \tag{47}$$

where

$$A_{ji} = \sum_{k=0}^{j} \binom{j}{k} \theta(i-k)\, [\delta_{0,\, i-k}\, A^{(j-k)} + \beta^{(j-k)} K_{0,\, i-k}] \tag{48}$$

$$i = 0, \ldots, m$$

where $\binom{j}{k}$ is the binomial coefficient.

In accordance with equation (39) it is easily shown that

$$\beta^{(k)} = \sum_{j=1}^{k} \binom{k}{j} \frac{\partial^j B}{\partial \alpha^j} R^{-1} \frac{\partial^{k-j} B^T}{\partial \alpha^{k-j}} \tag{49}$$

2. Simulation

Equation (12) can be used as the basis for simulation. By writing $\hat{A}$ as

$$\hat{A} = [\hat{A}_0^T\ \hat{A}_1^T\ .\ .\ .,\ \hat{A}_m^T]^T \tag{50}$$

where

$$\hat{A}_i = [A_{i0},\ A_{i1},\ .\ .\ .,\ A_{im}],\ i = 0, 1,\ .\ .\ .,\ m \tag{51}$$

one can write equation (12) as the m+1 equations

$$\begin{aligned} \dot{x} &= Ax + Bu \\ \dot{v}_i &= \hat{A}_i X,\ i = 1, 2,\ .\ .\ .,\ m \end{aligned} \tag{52}$$

Further, from equations (13), (16) and (33) one can write

$$u = -R^{-1}B^T K_{00}x - R^{-1}B^T \sum_{i=1}^{m} K_{0i}\, v_i \tag{53}$$

Equations (52) and (53) lead to the simulation shown in Figure 1.

3. Illustrative Example

Consider a first-order system characterized by

$$\dot{x} = -a^2 x + u$$

It is desired to find the control u that minimizes the performance index

$$J = \int_0^\infty (x^2 + u^2 + s\,v^2)\,dt$$

where $v = \partial x/\partial a$.

For this example, the optimal control is given by $u = -k_{00}x - k_{01}v$. The feedback coefficients k_{00} and k_{01} are found by solving equation (14) for K. It is noted that B=1, R=1, Q=1, S=s and

$$\hat{A} = \begin{bmatrix} -a^2 & 0 \\ -2a & a^2 - k_{00} \end{bmatrix}, \quad K = \begin{bmatrix} k_{00} & k_{01} \\ k_{01} & k_{11} \end{bmatrix}$$

For the case that the nominal parameter value $a = a_0 = 1$ we obtain: for s = 1, $u = -0.65x + 0.178v$; for s = 10, $u = -1.33x + 0.855v$; for s = 100, $u = -2.81x + 3.13v$.

Figure 2 shows the response, for various values of a, of the optimal reference system; trajectory sensitivity is not considered, i.e., s = 0. Figures 3, 4 and 5 show the responses of the optimal system for s = 1, s = 10 and s = 100 respectively; each shows the responses for various values of a. These figures also show the trajectory sensitivity v generated by the nominal optimal system compared the computed values of trajectory sensitivity (dotted curve); the computation involves the taking the difference of two responses of the optimal system with plus and minus deviations of the parameter nominal value. As expected, the system responses deviate less from the nominal trajectory (note the smaller trajectory sensitivity) for larger values of s.

4. Acknowledgements

The authors gratefully acknowledge the support of the National Science Foundation (Grant No. GK 274) and NASA (Grant No. NsG-518). The authors also wish to express their thanks to Mr. Dan Curran who did all the digital and analog computer programming for the many numerical problems worked.

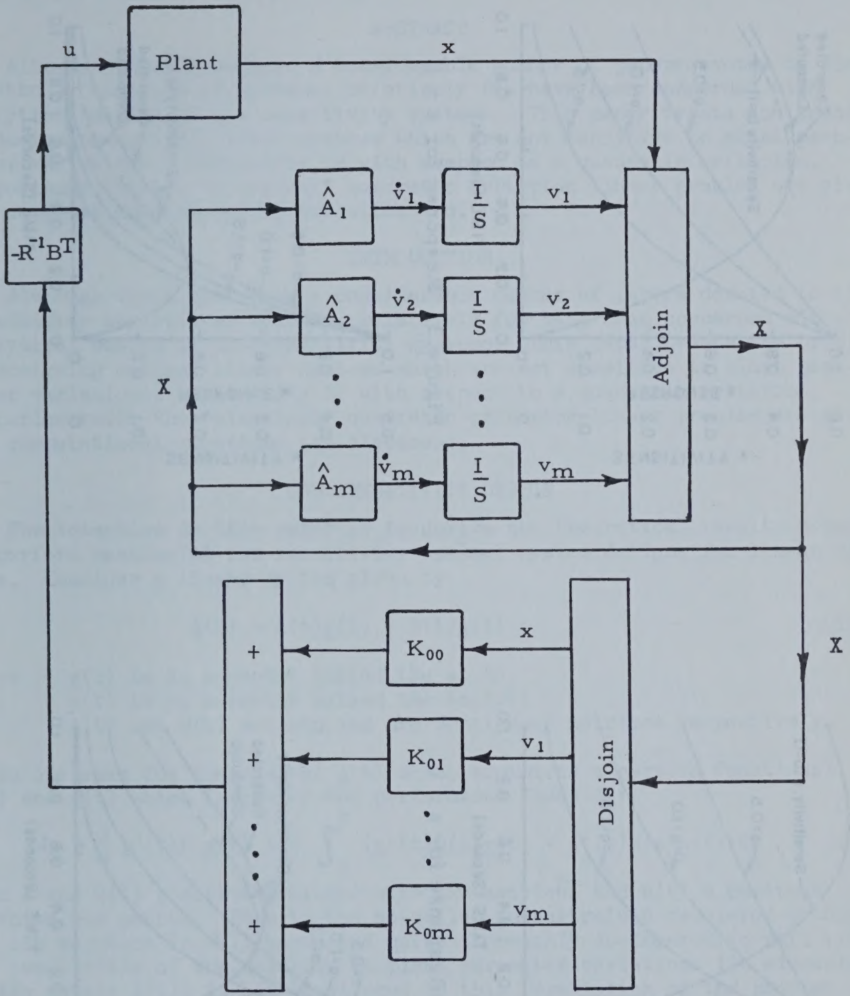

u
Plant
x
$-R^{-1}B^{T}$
$\hat{A}_1$
$\dot{v}_1$
$\frac{I}{S}$
v_1
$\hat{A}_2$
$\dot{v}_2$
$\frac{I}{S}$
v_2
X
$\hat{A}_m$
$\dot{v}_m$
$\frac{I}{S}$
v_m
Adjoin
X
X
+
K_{00}
x
+
K_{01}
v_1
Disjoin
+
K_{0m}
v_m

Figure 1. Simulation of Optimal Control

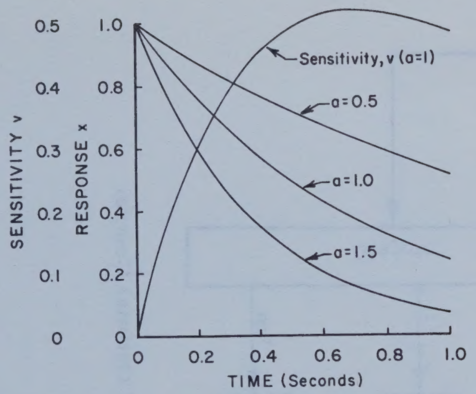


Figure 2. Response for s = 0

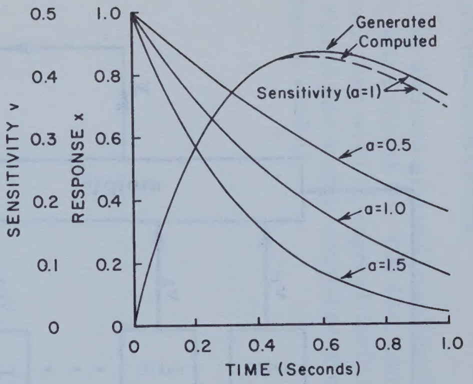


Figure 3. Response for s = 1

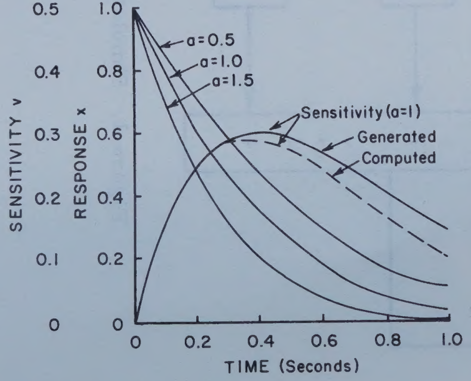


Figure 4. Response for s = 10

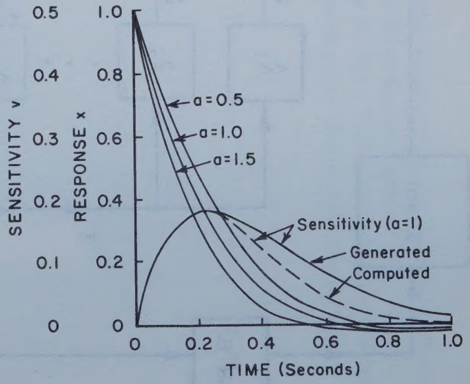


Figure 5. Response for s = 100

LOW SENSITIVITY DESIGN OF OPTIMAL LINEAR CONTROL SYSTEMS

STEPHEN J. KAHNE
University of Minnesota
Minneapolis, Minnesota 55455

ABSTRACT

Although there have been a considerable number of papers devoted to the sensitivity analysis of systems, relatively few have been concerned with analytical design of low sensitivity systems. This paper treats the problem of designing optimal linear systems which are not sensitive to plant parameter variations. Optimality is with respect to a quadratic criterion. Comparisons with the "classical" quadratic criterion linear problem are given and computational questions are discussed.

INTRODUCTION

Although there have been a considerable number of papers devoted to the sensitivity analysis of systems, relatively few have been concerned with analytical design of low sensitivity systems. This paper treats the problem of designing optimal linear systems which are not sensitive to plant parameter variations. Optimality is with respect to a quadratic criterion. Comparisons with the "classical" quadratic criterion linear problem are given and computational questions are discussed.

LOW-SENSITIVITY DESIGN

The intention in this paper is to derive the theoretical results necessary to perform meaningful low sensitivity optimal system designs for linear systems. Consider a linear system given by

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{u}(t) \tag{1}$$

where $\underline{x}(t)$ is an n-vector called the state
$\underline{u}(t)$ is an m-vector called the control
$\underline{A}(t)$ and $B(t)$ are nxn and nxm continuous matrices respectively.

Often one asks for the control $\underline{u}(t)$ which minimizes a certain functional $\underline{x}(t)$ and $\underline{u}(t)$ which is called the performance index J_1:

$$J_1 = \frac{1}{2}\, \underline{x}'(T) F\, \underline{x}(T) + \frac{1}{2} \int_0^T (\underline{x}'(t)Q(t)\underline{x}(t) + \underline{u}'(t)R(t)\underline{u}(t))dt \tag{2}$$

with F and Q(t) positive semi-definite nxn matrices and R(t) a positive definite mxm matrix. This is the so-called unconstrained regulator problem and its solution is well known and quite thoroughly documented in ref. [1]. The sensitivity of the solution to plant parameter variations (in elements of the matrix A(t)) is not considered in this formulation of the problem and often results in highly undesirable system designs from this standpoint.

In practical system design, sensitivity considerations play a fundamental role in the selection of controllers but in the literature of control theory there has been relatively little attention paid to the analytical aspects of this question. A theoretical basis for such studies is given here.

As is well known in the control literature[2] if $\underline{s}(t)$ is the sensitivity vector defined

$$\underline{s}(t) = \left.\frac{d\underline{x}(t)}{d\alpha}\right|_{\alpha = \alpha_{\text{nominal}}} \tag{3}$$

where α is a parameter of the system in the A matrix, then the equation describing the time behavior of $\underline{s}$ is

$$\text{neglecting } \left.\frac{\partial \underline{u}}{\partial \alpha}\right|_{\alpha=\alpha_{\text{nom.}}} \qquad \dot{\underline{s}}(t) = A(t)\underline{s}(t) + C(t)\underline{x}(t) \tag{4}$$

where C(t) is an nxn matrix of known structure. In fact C(t) has the elements noted:

$$[C(t)]_{ij} = 0 \qquad \text{if } [A]_{ij} \text{ does not contain } \alpha$$

$$[C(t)]_{ij} = \left.\frac{d[A]_{ij}}{d\alpha}\right|_{\alpha=\alpha_{\text{nominal}}} \qquad \text{if } [A]_{ij} \text{ does contain } \alpha$$

Moreover, for the situation we are considering in which only a single parameter (not necessarily a single element) in A(t) is subject to variation, the initial conditions on Eq. (4) are:

$$\underline{s}(o) = \underline{0} \tag{5}$$

At the end of the paper are comments about other problems closely related to the one examined here. The results derived here are readily applicable to those problems.

Having defined $\underline{s}(t)$ we may now construct a sensitivity design criterion

$$J_2 = \frac{1}{2}\int_o^T (\underline{s}'(t)\ \Sigma(t)\underline{s}(t))dt \tag{6}$$

which, provided Σ(t) is a positive semi-definite matrix, is a meaningful measure of the sensitivity of a trajectory $\underline{x}(t)$ to changes in a parameter α of nominal value α_{nominal}.

Now we may state the linear regulator problem for low sensitivity design. Problem: Select, from the set of bounded functions, a $\underline{u}(t)$ which minimizes (for T fixed)

$$J = \frac{1}{2}\ \underline{x}'(T)F\ \underline{x}(T) + \frac{1}{2}\int_o^T (\underline{x}'(t)Q(t)\underline{x}(t) + \underline{u}'(t)R(t)\underline{u}(t) + \underline{s}'(t)\Sigma(t)\underline{s}(t))dt \tag{7}$$

subject to the differential constraints

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{u}(t) \qquad \underline{x}(o) = \underline{x_o} \tag{8}$$

$$\dot{\underline{s}}(t) = A(t)\underline{s}(t) + C(t)\underline{x}(t) \qquad \underline{s}(o) = \underline{0} \tag{9}$$

Let us assume that F, Q(t) and Σ(t) are symmetric positive semi-definite matrices with Q(t) and Σ(t) continuous and that R(t) is continuous, symmetric, and positive definite. The fact that F is constant should be recalled from basic physical considerations.

OPTIMAL CONTROL LAW DERIVATION

Pontryagin's Maximum Principle will be used here without further justification or comment. The Hamiltonian for this problem is

$$\begin{aligned} H = &\frac{1}{2}\ \underline{x}'(t)Q(t)\underline{x}(t) + \frac{1}{2}\ \underline{u}'(t)R(t)\underline{u}(t) \\ &+ \frac{1}{2}\ \underline{s}'(t)\Sigma(t)\underline{s}(t) + \underline{p}'(t)A(t)\underline{x}(t) + \underline{p}'(t)B(t)\underline{u}(t) \\ &+ \underline{\lambda}'(t)A(t)\underline{s}(t) + \underline{\lambda}'(t)C(t)\underline{x}(t) \end{aligned} \tag{10}$$

and the canonical equations follow,

$$\dot{\underline{x}}(t) = \frac{\partial H}{\partial \underline{p}} = A(t)x(t) + B(t)\underline{u}(t) \tag{11}$$

$$\dot{\underline{p}}(t) = \frac{-\partial H}{\partial \underline{x}} = -Q'(t)\underline{x}(t) - A'(t)\underline{p}(t) - C'(t)\underline{\lambda}(t) \tag{12}$$

$$\dot{\underline{s}}(t) = \frac{\partial H}{\partial \underline{\lambda}} = A(t)\underline{s}(t) + C(t)\underline{x}(t) \tag{13}$$

$$\dot{\underline{\lambda}}(t) = \frac{-\partial H}{\partial \underline{s}} = -\Sigma'(t)\underline{s}(t) - A'(t)\underline{\lambda}(t) \tag{14}$$

A necessary condition for $\underline{u}(t)$ to be optimal is that

$$\frac{\partial H}{\partial \underline{u}} = \underline{0} = R'(t)\underline{u}(t) + B'(t)\underline{p}(t) \tag{15}$$

Finally, boundary conditions are given for each equation

$$(11): \quad \underline{x}(o) = \underline{x_o} \tag{16}$$

$$(12): \quad \underline{p}(T) = F\underline{x}(T) \tag{17}$$

$$(13): \quad \underline{s}(o) = \underline{0} \tag{18}$$

$$(14): \quad \underline{\lambda}(T) = \underline{0} \tag{19}$$

The reader is referred to ref. [1] for proofs of these statements.

From (15) we have

$$\underline{u}(t) = -R'^{-1}(t)B'(t)\underline{p}(t) \tag{20}$$

Using (20) in (11) we get

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) - B(t)R'^{-1}(t)B'(t)\underline{p}(t) \tag{21}$$

Then (12), (13), (14) and (21) can be combined to give

$$\begin{bmatrix} \dot{\underline{x}} \\ \dot{\underline{s}} \\ \dot{\underline{p}} \\ \dot{\underline{\lambda}} \end{bmatrix} = \left[\begin{array}{c|c|c|c} A(t) & 0 & -B(t)R'^{-1}(t)B'(t) & 0 \\ \hline C(t) & A(t) & 0 & 0 \\ \hline -Q'(t) & 0 & -A'(t) & -C'(t) \\ \hline 0 & -\Sigma'(t) & 0 & -A'(t) \end{array}\right] \begin{bmatrix} \underline{x} \\ \underline{s} \\ \underline{p} \\ \underline{\lambda} \end{bmatrix} \tag{22}$$

and boundary conditions for (22) are given in component form by (16) - (19). We may think of $\underline{x}$ and $\underline{s}$ as a super-state vector (dimension 2n)

$$\underline{z} = \left[\frac{\underline{x}}{\underline{s}}\right] \tag{23}$$

and $\underline{p}$ and $\underline{\lambda}$ as a super-adjoint vector (dimension 2n)

$$\psi = \left[\frac{\underline{p}}{\underline{\lambda}}\right] \tag{24}$$

and then write (22) as

$$\left[\frac{\dot{\underline{z}}}{\dot{\psi}}\right] = \Gamma(t) \left[\frac{\underline{z}}{\underline{\psi}}\right] \tag{25}$$

where $\Gamma(t)$ is clearly the (4n x 4n) system matrix of (22). If we next denote the transition matrix of (25) by $\Phi(t,t_o)$, then the solution to (25) may be written

$$\begin{bmatrix}\underline{z}(t)\\ \underline{\psi}(t)\end{bmatrix} = \Phi(t,t_o)\begin{bmatrix}\underline{z}(t_o)\\ \underline{\psi}(t_o)\end{bmatrix} \tag{26}$$

or, expressing $\Phi(t,t_o)$ in four components (each 2n x 2n)

$$\begin{bmatrix}\underline{z}(t)\\ \underline{\psi}(t)\end{bmatrix} \begin{matrix}=\\=\end{matrix} \begin{bmatrix}\varphi_{11}(t,t_o) & \varphi_{12}(t,t_o)\\ \varphi_{21}(t,t_o) & \varphi_{22}(t,t_o)\end{bmatrix}\begin{bmatrix}\underline{z}(t_o)\\ \underline{\psi}(t_o)\end{bmatrix} \tag{27}$$

If we next evaluate (27) at t = T, introduced in (2), and recall that $\underline{\psi}(T) = \begin{bmatrix}F\underline{x}(T)\\ \underline{0}\end{bmatrix}$ from (17) and (19) we get

$$\begin{aligned}\underline{z}(T) &= \varphi_{11}(T,t_o)\underline{z}(t_o) + \varphi_{12}(T,t_o)\underline{\psi}(t_o)\\ \underline{\psi}(T) &= \varphi_{21}(T,t_o)\underline{z}(t_o) + \varphi_{22}(T,t_o)\underline{\psi}(t_o) = \begin{bmatrix}F\ \underline{x}(T)\\ \underline{0}\end{bmatrix} = G\ \underline{z}(T)\end{aligned} \tag{28}$$

where $G = \left[\begin{array}{c:c} F & 0\\ \hdashline 0 & 0\end{array}\right]$

Since F is positive semi-definite, so is G. Realizing that (28) is true for any t_o, we call it t and deduce from (28) that

$$\underline{\psi}(t) = [\varphi_{22}(T,t) - G\varphi_{12}(T,t)]^{-1}[G\varphi_{11}(T,t) - \varphi_{21}(T,t)]z(t) \tag{29}$$

We write (29) as

$$\underline{\psi}(t) = K(t)\underline{z}(t) \tag{30}$$

provided the inverse in (29) exists. The proof of the existence of this inverse is in ref. [3]. Lets look at (29) when t = T. Then

$$\underline{\psi}(T) = [\varphi_{22}(T,T) - G\varphi_{12}(T,T)]^{-1}[G\varphi_{11}(T,T) - \varphi_{21}(T,T)]\underline{z}(T) \tag{31}$$

Since $\Phi(T,t)$ is a transition matrix we know that $\Phi(T,T) = I$, the 4n x 4n identity matrix. Thus $\varphi_{11}(T,T) = \varphi_{22}(T,T) = I$ and $\varphi_{21}(T,T) = \varphi_{12}(T,T) = 0$ the zero (2n x 2n) matrix. Thus from (31)

$$\underline{\psi}(T) = G\underline{z}(T) \tag{32}$$

so, from (30) and (32)

$$K(T) = G \tag{33}$$

The basic result between Eqs. (22) and (33) is that $\underline{\psi}(t) = K(t)\underline{z}(t)$ and $K(T) = G$. These results are discussed in more detail in ref. [1] for the interested reader.

PROPERTIES OF THE LINEAR FEEDBACK GAIN MATRIX

Let's examine (30) more closely

$$\underline{\psi}(t) = \begin{bmatrix}\underline{p}(t)\\ \underline{\lambda}(t)\end{bmatrix} = \begin{bmatrix}K_{11}(t) & K_{12}(t)\\ K_{21}(t) & K_{22}(t)\end{bmatrix}\begin{bmatrix}\underline{x}(t)\\ \underline{s}(t)\end{bmatrix} \tag{34}$$

$$K_{12}(T) = K_{22}(T) = K_{21}(T) = 0, \quad K_{11}(T) = F \tag{35}$$

Thus

$$\underline{p}(t) = K_{11}(t)\underline{x}(t) + K_{12}(t)\underline{s}(t) \tag{36}$$

and, from (20), the optimal control is

$$\underline{u}(t) = -R'^{-1}(t)B'(t)K_{11}(t)\underline{x}(t) - R'^{-1}(t)B'(t)K_{12}(t)\underline{s}(t) \tag{37}$$

The "classical" optimal regulator problem solution looked very much like this except that the second term was missing. Our next task will be to determine properties of $K_{11}(t)$ and $K_{12}(t)$ to permit their calculation and establish the nature of the controller implied by (37). To do this we again proceed as in ref. [1]. Differentiating (30) with respect to time gives

$$\dot{\underline{\psi}}(t) = \dot{K}(t)\underline{z}(t) + K(t)\dot{\underline{z}}(t) \tag{38}$$

From (22)

$$\dot{\underline{z}}(t) = L(t)\underline{z}(t) - M(t)\underline{\psi}(t) \tag{39}$$

$$\dot{\underline{\psi}}(t) = -N(t)\underline{z}(t) - P(t)\underline{\psi}(t) \tag{40}$$

where

$$L(t) = \left[\begin{array}{c|c} A(t) & 0 \\ \hline C(t) & A(t) \end{array}\right] \tag{41}$$

$$M(t) = \left[\begin{array}{c|c} B(t)R'^{-1}(t)B'(t) & 0 \\ \hline 0 & 0 \end{array}\right] \tag{42}$$

$$N(t) = \left[\begin{array}{c|c} Q'(t) & 0 \\ \hline 0 & \Sigma'(t) \end{array}\right] \tag{43}$$

$$P(t) = \left[\begin{array}{c|c} A'(t) & C'(t) \\ \hline 0 & A'(t) \end{array}\right] \tag{44}$$

all 2n x 2n matrices. Using (30) in (39) gives

$$\dot{\underline{z}}(t) = (L(t) - M(t)K(t))\underline{z}(t) \tag{45}$$

and (45) into (38) gives

$$\dot{\underline{\psi}}(t) = [\dot{K}(t) + K(t)L(t) - K(t)M(t)K(t)]\underline{z}(t) \tag{46}$$

Next put (30) into (40):

$$\dot{\underline{\psi}}(t) = -[N(t) + P(t)K(t)]\underline{z}(t) \tag{47}$$

(46) and (47) yield

$$[\dot{K}(t) + K(t)L(t) + P(t)K(t) - K(t)M(t)K(t) + N(t)]\underline{z}(t) = \underline{0} \tag{48}$$

(48) is to be true for all $t \epsilon [0,T]$. From (48) we deduce the following equation for $K(t)$

$$\dot{K}(t) + K(t)L(t) + P(t)K(t) - K(t)M(t)K(t) + N(t) = 0 \tag{49}$$

Next we take a more detailed look at (49) and omit all dependences on t for convenience:

$$\begin{bmatrix} \dot K_{11} & \dot K_{12} \\ \dot K_{21} & \dot K_{22} \end{bmatrix} + \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} A & 0 \\ C & A \end{bmatrix} + \begin{bmatrix} A' & C' \\ 0 & A' \end{bmatrix} \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} -$$

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} B\,R'^{-1}B' & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} + \begin{bmatrix} Q' & 0 \\ 0 & \Sigma' \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad (50)$$

or

$$\begin{bmatrix} \dot K_{11} & \dot K_{12} \\ \dot K_{21} & \dot K_{22} \end{bmatrix} + \begin{bmatrix} K_{11}A + K_{12}C & K_{12}A \\ K_{21}A + K_{22}C & K_{22}A \end{bmatrix}$$

$$+ \begin{bmatrix} A'K_{11} + C'K_{21} & A'K_{12} + C'K_{22} \\ A'K_{21} & A'K_{22} \end{bmatrix}$$

$$- \begin{bmatrix} K_{11}B\,R'^{-1}B'K_{11} & K_{11}B'R'^{-1}BK_{12} \\ K_{21}B\,R'^{-1}B'K_{11} & K_{21}B'R'^{-1}BK_{12} \end{bmatrix} + \begin{bmatrix} Q' & 0 \\ 0 & \Sigma' \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad (51)$$

(51) yields the following four nxn matrix equations:

$$\dot K_{11} + K_{11}A + K_{12}C + A'K_{11} + C'K_{21} - K_{11}B\,R'^{-1}B'K_{11} + Q' = 0 \qquad (52)$$

$$\dot K_{12} + K_{12}A + A'K_{12} + C'K_{22} - K_{11}B\,R'^{-1}B'K_{12} = 0 \qquad (53)$$

$$\dot K_{21} + K_{21}A + K_{22}C + A'K_{21} - K_{21}B\,R'^{-1}B'K_{11} = 0 \qquad (54)$$

$$\dot K_{22} + K_{22}A + A'K_{22} - K_{21}B\,R'^{-1}B'K_{12} + \Sigma' = 0 \qquad (55)$$

with boundary conditions given in (35).

It is interesting to compare this result with the "classical" results of optimizing linear systems with respect to quadratic criteria. In the latter case, C = 0 and the equation for the feedback matrix is just given by (52). Note that the terms coupling (52) with (53) - (55) drop out in this case. The boundary condition in that case is K(T) = F which is just the boundary condition on $K_{11}(T)$ in (35). The results of "classical" theory and the present paper thus agree.

Now, if K(t) has no symmetry properties, it appears that it will be necessary to solve 2n x 2n or $4n^2$ scalar equations. So, let us examine K(t) for possible symmetry. Take the transpose of (49):

$$\dot K'(t) + L'(t)K'(t) + K'(t)P'(t) - K'(t)M'(t)K'(t) + N'(t) = 0$$

For any matrix it is true that

$$\left[\frac{d}{dt} K(t)\right]' = \frac{d}{dt} K'(t) \qquad (56)$$

Since $L(t) = \begin{bmatrix} A(t) & 0 \\ C(t) & A(t) \end{bmatrix}$

and $P(t) = \begin{bmatrix} A'(t) & C'(t) \\ 0 & A'(t) \end{bmatrix}$

and since $A(t) = A'(t)$, $C(t) = C'(t)$ it is clear that

$$(K(t)L(t))' = L'(t)K'(t) = P(t)K'(t) \tag{57}$$

and $$(P(t)K(t))' = K'(t)P'(t) = K'(t)L(t) \tag{58}$$

Also, since $$M(t) = \begin{bmatrix} B(t)R'^{-1}(t)B'(t) & 0 \\ 0 & 0 \end{bmatrix}$$

$$M(t) = M'(t)$$

and $$(K(t)M(t)K(t))' = K'(t)M(t)K'(t) \tag{59}$$

Finally, $$N(t) = \begin{bmatrix} Q'(t) & 0 \\ 0 & \Sigma'(t) \end{bmatrix}$$

is also symmetric so $N(t) = N'(t)$. Thus the transpose of (49) may be written

$$\dot{K}'(t) + P(t)K'(t) + K'(t)L(t) - K'(t)M(t)K'(t) + N(t) = 0 \tag{60}$$

(60) and (49) are identical equations for $K(t)$ and $K'(t)$ and since their boundary condition $K(T) = G$ is also symmetric we have

$$K'(T) = G \tag{61}$$

Thus $K(t)$ and $K'(t)$ satisfy the same equation and same boundary condition and from the theory of ordinary linear differential equations

$$K(t) = K'(t) \tag{62}$$

meaning that $K(t)$ is symmetric. This means that $K_{11}(t)$ and $K_{22}(t)$ are symmetric and $K_{12}(t) = K_{21}(t)$. Thus we are only required to solve $n(2n + 1)$ simultaneous differential equations. The following table shows how many equations may be expected in general. Column A is the dimension of the state vector. Column B is the number of equations required for the linear system optimization with quadratic performance criterion and no sensitivity considerations. Column C shows the number of equations to be solved when a low-sensitivity optimal design is required. Column D lists the number of equations in the latter case if the symmetry properties derived above were not used.

A	B	C	D
n	$\frac{n}{2}(n+1)$	$n(2n+1)$	$4n^2$
1	1	3	4
2	3	10	16
3	6	21	36
4	10	36	64
5	15	55	100
6	21	78	144

Since the $C(t)$ and $\Sigma(t)$ matrices are likely to be quite sparce, many of the $n(2n+1)$ equations may be trivial and the computations required will be greatly simplified. However, in the case of a study of effects on the system due to several parameter changes the number will rise again. Before one can make conclusive statements about computational problems, considerable experience with sample calculations will be required.

SUMMARY OF RESULTS

It is of interest to compare solutions of the ordinary optimal linear system problem and the low sensitivity optimal linear system problem. In both cases the usual limitations and assumptions apply.

I. Optimal Control of a Linear System with Respect to a Quadratic Criterion

System: $$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{u}(t) \tag{63}$$

Initial Condition: $$\underline{x}(o) = \underline{x_o} \tag{64}$$

Performance Criterion: $$J = \underline{x}'(T)F\underline{x}(T) + \frac{1}{2}\int_0^T (\underline{x}'(t)Q(t)\underline{x}(t)+\underline{u}'(t)R(t)\underline{u}(t))dt \tag{65}$$

Optimal Control: $$\underline{u}(t) = -R'^{-1}(t)B'(t)K(t)\underline{x}(t) \tag{66}$$

Gain Matrix: $$\dot{K}(t) + K(t)A(t) + A'(t)K(t) - K(t)B'(t)R'^{-1}(t)B'(t)K(t) + Q(t) = 0 \qquad K(T) = F \tag{67}$$

II. Optimal Low Sensitivity Control of a Linear System with Respect to a Quadratic Criterion

System: $$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{u}(t) \tag{68}$$

Initial Condition: $$\underline{x}(o) = \underline{x_o} \tag{69}$$

Sensitivity of (68) with respect to a parameter in $A(t)$:

$$\dot{\underline{s}}(t) = A(t)\underline{s}(t) + C(t)\underline{x}(t) \tag{70}$$

Initial Condition: $$\underline{s}(o) = \underline{0} \tag{71}$$

Performance Criterion: $$J = \underline{x}'(T)F\underline{x}(T) + \frac{1}{2}\int_0^T (\underline{x}'(t)Q(t)\underline{x}(t) + \underline{u}'(t)R(t)\underline{u}(t) + \underline{s}'(t)\Sigma(t)\underline{s}(t))dt \tag{72}$$

Optimal Control: $$\underline{u}(t) = -R'^{-1}(t)B'(t)K_{11}(t)\underline{x}(t) - R'^{-1}(t)B'(t)K_{12}(t)\underline{s}(t) \tag{73}$$

Gain Matrices: See (52) to (55) and (35).

Several interesting observations may be made concerning these results.

1. Whereas, in problem I it is meaningless to consider the case where $F = Q(t) = 0$, that is not the case for problem II. This is true because the cost of the state deviations is reflected in the $\underline{s}'(t)\Sigma(t)\underline{s}(t)$ term in (72).

2. It follows directly from well known results that in the limit as $T\to\infty$, if all matrices in the system and performance criterion are constant, that K is constant for problem I and is given by (67) with $\dot{K}(t) = 0$. For problem II, K_{11}, K_{12}, K_{21} and K_{22} are constants given by (52) to (55) with all derivatives equal to zero.

3. The structure of the low sensitivity optimal control system is conveniently shown by Fig. 1.

FURTHER QUESTIONS

To determine the usefulness of these results, considerable numerical calculation will be needed. It is clear that the amount of effort will be somewhat greater than is required when no sensitivity terms are included in the performance criterion. However, from the standpoint of engineering design, an optimal controller which does not provide low sensitivity to plant parameter changes is of questionable value. A related problem which must be examined is the problem of several parameters in A being subject to variation or uncertainty. This appears to be more of a notational than a theoretical question. Techniques of the present paper may be used to examine the question one parameter at a time. Probably trade-offs between competing designs would be necessary to determine a satisfactory final design. Again numerical experimentation may be the best answer here.

REFERENCES

1. Athans, M. and Falb, P., Optimal Control, McGraw-Hill, 1966.

2. Tuel, W., Lee, I. and DeRusso, P., "Synthesis of Optimal Control Systems with Sensivitiy Constraints," Proc. Third IFAC Congress, London, 1966.

3. Kalman, R. E., "Contributions to the Theory of Optimal Control," Bol. Soc. Math. Mexico, pp. 102-119 (1960).

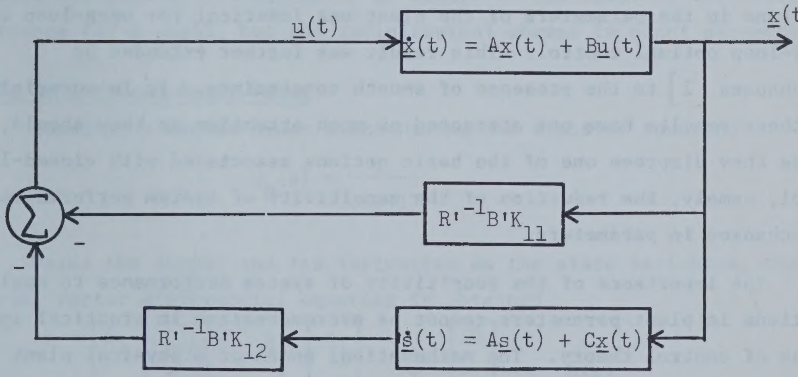

$\underline{u}(t)$
$\dot{\underline{x}}(t) = A\underline{x}(t) + B\underline{u}(t)$
$\underline{x}(t)$
Σ
$-$
$-$
$R'^{-1}B'K_{11}$
$R'^{-1}B'K_{12}$
$\dot{\underline{s}}(t) = A\underline{s}(t) + C\underline{x}(t)$

Figure 1 Structure of Low Sensitivity Optimal Control System

SENSITIVITY OF OPTIMAL CONTROL SYSTEMS

N. K. SINHA and SATYA RATNAM ATLURI
Department of Electrical Engineering
McMaster University
Hamilton, Ontario, Canada

ABSTRACT

The sensitivity of the index of performances to variations in plant parameters is considered for open-loop optimal control as well as closed-loop optimal control with one or two degrees of freedom. It has been shown that, in general, the sensitivity of the index of performances to slight variations in the location of a pole of the plant, is smaller for closed-loop optimal than for the open-loop case. An example of a simple second-order plant has been used to further illustrate the theory, which is contrary to some recently published work.

Introduction

In a recent paper, Pagurek [1] obtained an extraordinary result on linear optimal control systems with quadratic performance criteria. He proved that the sensitivity of the index of performance with respect to variations in the parameters of the plant was identical for open-loop and closed-loop optimal control. This result was further extended by Witsenhausen [2] to the presence of smooth constraints. It is surprising that these results have not attracted as much attention as they should, because they disprove one of the basic notions associated with closed-loop control, namely, the reduction of the sensitivity of system performance to small changes in parameters.

The importance of the sensitivity of system performance to small variations in plant parameters cannot be overemphasized in practical applications of control theory. The mathematical model of a physical plant represents it only approximately, and small differences between the actual plant and the model are bound to exist. Aging, environmental effects, unavoidable inaccuracies due to tolerances in manufacture, and other unforeseen discrepancies, make it necessary for the system designer to reduce the sensitivity to small variations in the parameters. Feedback, therefore, has proved very useful to the designer in meeting stringent specifications economically with mass-produced components. On the other hand, it is highly desirable that the performance of the system be optimal, and it may often be worth while to pay the price of optimality, in terms of additional components. But Pagurek's result comes as a great blow to the designer, because it implies that if the system is to be optimized, only highly accurate components must be used, and feedback can no longer be utilized to reduce the sensitivity of performance to small changes in plant parameters.

The object of this paper is to investigate if the picture is really as gloomy as it appears. It appears that both Pagurek and Witsenhausen have used a definition of sensitivity which is valid only for an infinitesimal change in plant parameters, whereas in practical cases there is always a small but finite variation. This may, therefore, be the reason for their strange result. As a matter of fact, Witsenhausen [2] points out that closing the loop has no first-order effect, thereby implying that there may be higher-order effects which may not be negligible in the practical case of small, but not infinitesimal, changes in plant parameters. Moreover, it may also be noted that both the papers, referred to above, are of a rather abstract nature, and not even an example has been considered in either of them.

It was, therefore, decided to start by calculating the sensitivities for open-loop and closed-loop control for a particular example, using the practical definition of sensitivity as change in the value of the index of performance for a small, but not infinitesimal change in plant parameters.

Example of a Second-Order Plant

Consider a second-order plant having the transfer function

$$G_p(s) = \frac{1}{s(s+1)} \qquad (1)$$

Using the output and its derivative as the state variables, the following vector differential equation is obtained

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) \qquad (2)$$

where $x_1(t)$ is the output, $x_2(t)$ its first derivative, and $u(t)$ the input to the plant.

Given the boundary conditions

$$\left.\begin{aligned} \underline{x}(0) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \text{and} \qquad \underline{x}(t_f) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{aligned}\right\} \qquad (3)$$

it is required to find the optimal control input, $u^*(t)$, which would minimize the performance index:

$$J = \int_0^{t_f} x_1^2 \, dt \tag{4}$$

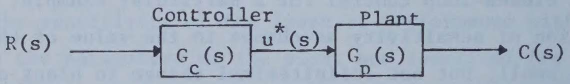
subject to the control energy input constraint

$$\int_0^{t_f} u^2 \, dt < 5 \tag{5}$$

Using Pontryagin's Maximum Principle, the optimum control input is obtained as

$$\begin{aligned} u^*(t) = {} & e^{-1.35t} (3.104 \cos 1.15t - 0.9517 \sin 1.15t) \\ & -e^{1.35t} (0.006777 \cos 1.15t - 0.00449 \sin 1.15t) \end{aligned} \tag{6}$$

The optimal open-loop linear control system is obtained by designing a controller which would convert a unit step input to $u^*(t)$, obtained in equation (6). The resulting system is shown in Figure 1.



$$G_p(s) = \frac{1}{s(s+1)}$$

$$G_c(s) = \frac{s(3.0972s^3 - 5.388s^2 + 1.6866s + 9.471)}{(s^4 - 0.998s^2 + 9.897)}$$

$$R(s) = \frac{1}{s}$$

FIGURE 1. Optimal Open-loop Control System

The optimal closed-loop system is obtained when part of the output is fed back to the input. It was decided to use a two-degree-of-freedom closed-loop system, as shown in Figure 2.

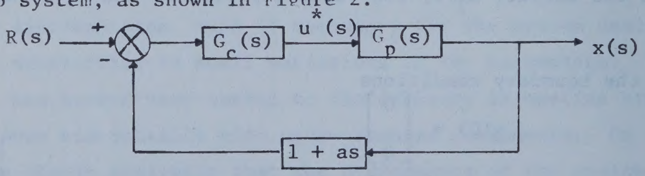


FIGURE 2. Two-Degree-of-Freedom Closed-Loop Optimal Control System

For different values of a, including zero, the transfer function of the controller that will cause $u^*(t)$ to be applied to the plant was calculated. For example, with a = 1,

$$G_c(s) = \frac{s(3.0972s^3 - 5.388s^2 + 1.6866s + 9.471)}{(s^4 - 3.0972s^3 + 4.390s^2 - 1.6866s + 0.421)} \tag{7}$$

and with a = 0,

$$G_c(s) = \frac{s(s+1)(3.0972s^3 - 5.388s^2 + 1.6866s + 9.471)}{(s^5 + s^4 - 4.0952s^3 + 4.390s^2 + 8.2104s + 0.426)} \tag{8}$$

The performance of index for each case is the same, and is computed as

$$J^* = 1.9094351 \tag{9}$$

Now, consider a change in the location of the pole of the plant from -1 to -1.01. In the case of the open-loop optimal system, the input to the plant is still unchanged, but the response, x_1, will now undergo a change. This is readily calculated, and the new performance index is obtained as

$$J' = 2.289334 \tag{10}$$

The sensitivity of the optimal open-loop system is, therefore, given by

$$S_o = \frac{\Delta J}{\Delta \alpha} = \frac{0.3194983}{0.01} = 31.94983 \tag{11}$$

In the case of the closed-loop optimal systems, however, a change in plant parameters also affects the input to the plant through the feedback loop. Calculating the performance indices for a = 2, and a = 0, the sensitivity, is obtained as,

$$S_c = 23.91260 \tag{12}$$

for a = 1, and

$$S_c' = 29.6772 \tag{13}$$

for a = 0.

Thus, it will be seen for this example, that the sensitivity of the index of performance to variations in the parameters of the plant is smaller for the closed-loop optimal system than for the open-loop optimal system. Moreover, the sensitivity can be further reduced by a suitable choice of a two-degree-of-freedom closed-loop optimal system, as seen with the case when a = 1.

A GENERAL EXPRESSION FOR THE SENSITIVITY OF OPTIMAL STATE-REGULATOR SYSTEMS

Consider a linear system, defined by the following equations:

$$\left.\begin{aligned} &\dot{\underline{x}} = A_o \underline{x} + B u \\ \text{and}\quad & \\ &\underline{x}(t_o) = \underline{x}_o \end{aligned}\right\} \qquad (14)$$

where $\underline{x}$ is an n-dimensional state-vector, $\underline{u}$ is an r-dimensional control input vector, A_o is an nxn matrix, and B is an nxr matrix. A_o is the nominal value of the matrix A, defined by the equation

$$A = f(\underline{a}) \qquad (15)$$

where $\underline{a}$ is an m-dimensional vector representing plant parameters, with the nominal value $A_o = f(\underline{a}_o)$.

The index of performance is defined as

$$V = \frac{1}{2}\int_{t_o}^{T} \left[\langle \underline{x}, Q\underline{x} \rangle + \langle \underline{u}, R\underline{u} \rangle \right] dt + \frac{1}{2} \langle \underline{x}(t_f), K\underline{x}(t_f) \rangle \qquad (16)$$

It follows from Kalman [3] that the optimal control input is given by

$$\underline{u}^*(t) = -R^{-1}B^T M_o(t,t_f)\, \underline{x}(t) \qquad (17)$$

where $M_o(t, t_f)$ is the solution of the matrix Riccati equation,

$$\left.\begin{aligned} &\frac{dM_o}{dt} + A_o^T M_o + M_o A_o - M_o B R^{-1} B^T M_o + Q_o = 0 \\ \text{and}\quad & \\ &M_o(t_f, t_f) = K \end{aligned}\right\} \qquad (18)$$

In the case of time-invariant linear systems, the matrices A_o, B, Q and R may all be time-invariant. For such a system, M_o is also time-invariant, so that $\frac{dM_o}{dt} = 0$ in equation (18)

The optimal closed-loop system follows from equations (14) and (17), and its sturcture is shown in Figure 3.

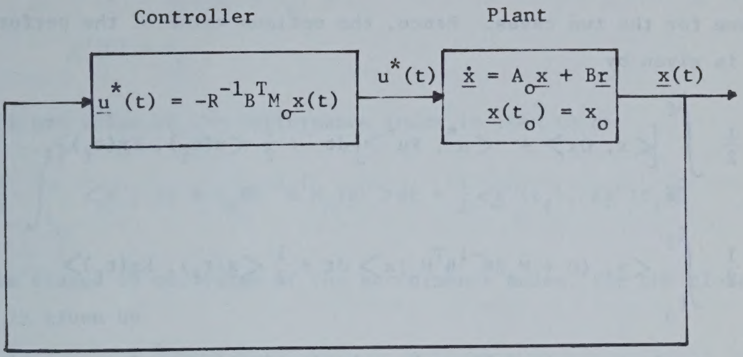

Controller
Plant
$\underline{u}^*(t) = -R^{-1}B^TM_o\underline{x}(t)$
$u^*(t)$
$\dot{\underline{x}} = A_o\underline{x} + B\underline{r}$
$\underline{x}(t_o) = x_o$
$\underline{x}(t)$

FIGURE 3: Optimal Closed-loop System

The optimal open-loop system is obtained by substituting for $\underline{u}^*(t)$ in equation (14) from (17),

$$\dot{\underline{x}} = A_o\underline{x}(t) - BR^{-1}B^TM_o(t, t_f)\,\underline{x}(t)$$

or,

$$\dot{\underline{x}} = P\,\underline{x}(t) \tag{19}$$

where

$$P = A_o - BR^{-1}B^TM_o(t, t_f) \tag{20}$$

The solution of equation (19) is

$$\underline{x}(t) = \Phi(t, t_o)\,\underline{x_o} \tag{21}$$

where $\Phi(t, t_o)$ is the transition matrix for the system defined by P, and for the time-invariant case

$$\Phi(t, t_o) = e^{P(t - t_o)} \tag{22}$$

Hence, the structure of the open-loop optimal system is as shown below:

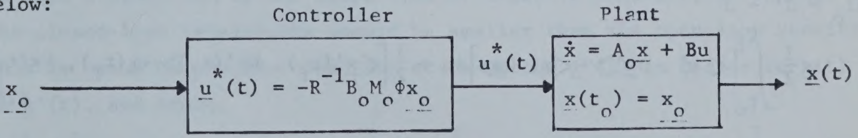

Controller
Plant
$\underline{x_o}$
$\underline{u}^*(t) = -R^{-1}B_oM_o\Phi\underline{x_o}$
$u^*(t)$
$\dot{x} = A_o\underline{x} + B\underline{u}$
$\underline{x}(t_o) = \underline{x_o}$
$\underline{x}(t)$

FIGURE 4: Optimum Open-loop Control System

Since, for the nominal values of the plant parameters, the same values of $\underline{u}^*(t)$ and $\underline{x}(t)$ are obtained for both the open-loop and closed-loop optimal systems, it follows that the value of the performance index will be the same for the two cases. Hence, the optimum value of the performance index is given by

$$J^* = \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x}, Q\underline{x}\rangle + \langle \underline{u}^*, R\underline{u}^*\rangle\right] dt + \frac{1}{2}\langle \underline{x}(t_f), K\underline{x}(t_f)\rangle$$

$$= \frac{1}{2}\int_{t_o}^{t_f}\langle \underline{x}, (Q + M_o BR^{-1}B^T M_o)\underline{x}\rangle \, dt + \frac{1}{2}\langle \underline{x}(t_f), K\underline{x}(t_f)\rangle \tag{23}$$

The last result follows by substituting for $\underline{u}^*$ from equation (17), and recalling that M_o is a symmetric matrix [4].

Now, consider a small variation in the plant parameters from $\underline{a}_o$ to $\underline{a}$, changing the matrix A_o to A.

In the case of the open-loop system, the resulting input to the plant will still be $\underline{u}^*(t)$, but the state vector $\underline{x}(t)$ would be different, say, $\underline{x}'(t)$ defined by the vector differential equation

$$\begin{aligned} \dot{\underline{x}} &= A\underline{x}' + B\underline{u}^* \\ \underline{x}'(0) &= \underline{x}_o \end{aligned} \tag{24}$$

The new performance of index, for the open-loop system is, therefore, given by

$$J' = \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x}', Q\underline{x}'\rangle + \langle \underline{u}^*, R\underline{u}^*\rangle\right] dt + \frac{1}{2}\langle \underline{x}'(t_f), K\underline{x}'(t_f)\rangle$$

$$= \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x}', Q\underline{x}'\rangle + \langle \underline{x}, M_o BR^{-1}B^T M_o \underline{x}\rangle\right]dt + \frac{1}{2}\langle \underline{x}'(t_f), K\underline{x}'(t_f)\rangle \tag{25}$$

The change in the value of the performance index, for the open-loop case, is given by

$$\begin{aligned} \Delta J_o &= J' - J^* \\ &= \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x}', Q\underline{x}'\rangle - \langle \underline{x}, Q\underline{x}\rangle\right]dt + \frac{1}{2}\left[\langle \underline{x}'(t_f), K\underline{x}'(t_f)\rangle - \langle \underline{x}(t_f), K\underline{x}(t_f)\rangle\right] \end{aligned} \tag{26}$$

In the case of the closed-loop system, a change in plant parameters also effects the input u(t), which will no longer be $u^*(t)$. Let the new state vector be $\underline{x}''(t)$, which is defined by

$$\dot{\underline{x}}'' = (A - BR^{-1}B^T M_o)\,\underline{x}''$$

$$\underline{x}''(0) = \underline{x}_o \tag{27}$$

The new value of the performance index is seen to be

$$J'' = \frac{1}{2}\int_{t_o}^{t_f} \langle \underline{x}'',\ (Q + M_o BR^{-1}B^T M_o)\underline{x}''\rangle dt + \frac{1}{2}\langle \underline{x}''(t_f),\ K\underline{x}''(t_f)\rangle \tag{28}$$

The change in the value of the performance index, for the closed-loop case is given by

$$\begin{aligned}\Delta J_c &= J'' - J^* \\ &= \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x}'',\ (Q + M_o BR^{-1}B^T M_o)\ \underline{x}''\rangle - \langle \underline{x},\ (Q + M_o BR^{-1}B^T M_o)\underline{x}\rangle\right] dt \\ &\quad + \frac{1}{2}\left[\langle \underline{x}''(t_f),\ K\underline{x}''(t_f)\rangle - \langle \underline{x}(t_f),\ K\underline{x}(t_f)\rangle\right]\end{aligned} \tag{29}$$

Subtracting equation (29) from (26), we have

$$\begin{aligned}\Delta J_o - \Delta J_c &= \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x}',\ Q\underline{x}'\rangle - \langle \underline{x}'',\ Q\underline{x}''\rangle\right] dt \\ &\quad + \frac{1}{2}\int_{t_o}^{t_f}\left[\langle \underline{x},\ M_o BR^{-1}B^T M_o \underline{x}\rangle - \langle \underline{x}'',\ M_o BR^{-1}B^T M_o \underline{x}''\rangle\right] dt \\ &\quad + \frac{1}{2}\left[\langle \underline{x}'(t_f),\ K\underline{x}'(t_f)\rangle - \langle \underline{x}''(t_f),\ K\underline{x}''(t_f)\rangle\right]\end{aligned} \tag{30}$$

Examining equation (30), although the last term would be negligibly small for any stable regulator, the first two terms would seldom be zero, except when $\underline{x}(t)$, $\underline{x}'(t)$ and $\underline{x}''(t)$ are identical for all values of time. Moreover, in the particular case where the performance index is a quadratic form of the state vector only, and independent of the control input vector, the second term would also vanish; thus making the two sensitivities inequal. In such a case, due to the properties of negative feedback, it appears that the closed-loop sensitivity should be smaller than the open-loop sensitivity. This is based on the fact that the state vector $\underline{x}''(t)$ is closer to $\underline{x}(t)$ than is $\underline{x}'(t)$, and hence,

$$\left[\langle \underline{x}',\ Q\underline{x}'\rangle - \langle \underline{x},\ Q\underline{x}\rangle\right] > \left[\langle \underline{x}'',\ Q\underline{x}''\rangle - \langle \underline{x},\ Q\underline{x}\rangle\right] \tag{31}$$

Conclusion

It has been shown that the sensitivity of optimal control systems to plant parameter variations are smaller for closed-loop systems than for open-loop systems if quadratic performance criteria, based on state vectors, are used. It appears that results to the contrary have been obtained by others [1, 2] due to reason that they have defined the sensitivity differently, and regarded it as a first-order effect. Thus, their results are valid only for infinitesimal changes in plant parameters, and not for small, but finite changes, which occur in practice.

It is, indeed, quite reassuring to find that feedback does reduce sensitivity to plant parameters, even for optimal systems.

Acknowledgements

This work was supported by a research grant from the National Research Council of Canada, and formed part of Mr. Atluri's Master's Thesis. The authors are also thankful to Miss Marie Kaneary for typing this paper.

References

1 B. Pagurek: "Sensitivity of the performance of optimal control systems to plant parameter variations", IEEE Trans. on Auto. Control, Vol. AC-10, April, 1965, pp 178-180.

2 H. Witsenhausen: "On the sensitivity of optimal control systems", IEEE Trans. on Auto. Control, vol. AC-10, October, 1965, pp 495-496

3 R. E. Kalman: "Contributions to the theory of optimal control", Bol. soc. mat., Mexicana, 1960, pp 102-119

4 M. Athans and P. L. Falb: "Optimal Controls", McGraw-Hill Book Company, Inc., 1966.

A DESIGN TECHNIQUE FOR IMPROVING SYSTEM SENSITIVITY *

WILLIAM A. PORTER
System Engineering Laboratory, Department of Electrical Engineering,
The University of Michigan, Ann Arbor, Michigan

ABSTRACT

In this paper a linear plant with variable characteristics is considered. It is assumed that the plant is driven by an open loop controller in an optimal manner. For a particular task and a nominal plant characteristic the controller is assumed to have a surplus capacity (in the optimal sense). This surplus capacity is used to minimize the effects of flucuations in the plant characteristics about the nominal.

INTRODUCTION

The problem of system sensitivity can take many forms depending on the disturbances considered and the design constraints in effect. Several authors; Pagurek [1], Witsenhausen [2], Holtzman and Horing [3] and Porter [4] have considered sensitivity problems in optimal systems (or systems under optimal control). Rohrer and Sobral [5], and Porter [6] have considered the problem of defining a sensitivity measure. Perkins and Cruz [7], and Porter [8] have investigated the reduction of sensitivity in multivariate systems.

The present paper deals with a mixture of sensitivity and optimality. The particular problems considered here, however, do not seem to have been dealt with in any of the earlier analyses. In this paper, linear systems under minimum energy control are considered. It is assumed that, for a nominal plant, the controller capacity exceeds the control requirements. The investigation deals with the question: How can the surplus capacity be used to minimize the effects of fluxuations in the plant characteristics?

1. A Direct Sensitivity Minimization Problem

It is well known that minimum energy problems in linear systems have a natural identification in terms of linear transformations between appropriate Hilbert spaces (see for instance [9, chap. 4]). Throughout the present analysis, H_1 and H_2 will denote Hilbert spaces while the bounded linear transformation $T:H_1 \rightarrow H_2$ represents the nominal system map. To denote the system disturbance the symbol δT, which is also bounded and linear, will be used. The informal problem given above may be sharpened as follows. Let $\xi \in H_2$ represent the system task. Suppose that ξ has a preimage under T with less than the available energy. How can the surplus energy be used to minimize the effects on δT.

This problem may still be given several diverse analytic interpretations. In the following two specific cases which reflect different approaches to the basic problem are considered. The first and most direct of the two cases is the following:

---

[*] This research was supported by the U.S. Army Research Office-Durham under contract DA-31-124 ARD-D-391 and by the U.S. Air Force under contract AF 30 (602) 3546.

Problem 1. For fixed $\xi \in H_2$ such that $||T^{\dagger}\xi|| < k$ find a $u \in T^{-1}(\xi)$ which satisfies $||u|| < k$ while minimizing $||\delta Tu||$.

In this problem, $T^{-1}(\xi)$ denotes the set of all preimages of ξ under T and $T^{\dagger}$ denotes the pseudo inverse of T (see [9, chap. 4]). First it is noted that if $u_{\xi} = T^{\dagger}\xi$ happens to be in the null space of δT this control is a solution to the problem. We shall assume that this uninteresting case does not occur. Secondly if $u \in H_1$ is a solution, then it may be decomposed in the form

$$u = v + u_{\xi}, \quad v \in N(T) \text{ and } u_{\xi} = T^{\dagger}\xi \in N(T)^{\perp}$$

where N(T) is the null space of T and $N(T)^{\perp}$ its orthogonal complement. This decomposition follows easily from the fact that N(T) is closed and that $T^{\dagger}\xi$ is the unique preimage of ξ in $N(T)^{\perp}$. Since

$$||u||^2 = ||v||^2 + ||u_{\xi}||^2$$

it follows that an equivalent statement of Problem 1 is: Find an element $v \in N(T)$ which minimizes $||\delta Tu_{\xi} + \delta Tv||$ while satisfying $||v||^2 \leq k^2 - ||u_{\xi}||^2$.

Let P denote the orthogonal projection of H_1 onto N(T). Then if v is a solution of Problem 1 clearly

$$\delta TPv = \delta Tv.$$

Moreover if ω is any element satisfying $||\omega|| \leq k^2 - ||u_{\xi}||^2$ and minimizing $||\delta Tu_{\xi} + \delta TP\omega||$ then $v = P\omega$ is a solution to Problem 1. Thus by introducing the transformation $V = \delta TP$ the constraint $v \in N(T)$ will automatically be satisfied.

Now let $\overline{R(V)}$ denote the closure of the range of V. If ξ' is the orthogonal projection of ξ on $\overline{R(V)}$ and if $\xi' \in R(V)$ then $V^{-1}(\xi')$ is nonempty. If the minimum element of this set has norm $\leq k^2 - ||u_{\xi}||^2$ then a solution to Problem 1 is obtained by choosing v to be this minimum element. If the norm of this minimum element exceeds $k^2 - ||u_{\xi}||^2$ or if $\xi' \notin R(V)$ then any solution to the problem must have $||v|| = k^2 - ||u_{\xi}||^2$. It is this last case which merits further study.

If $\delta\xi = \delta Tu_{\xi} = \delta T\, T^{\dagger}\xi$ and ϵ is a scalar such that $k^2 = (1+\epsilon)||T^{\dagger}\xi||^2$ then it suffices to minimize the functional; $||\delta\xi + Vv||$, subject to the constraint $||v||^2 = \epsilon\, ||T^{\dagger}\xi||^2$. The Lagrange multiplier technique will be used. Forming the functional

$$J(v) = ||\delta\xi + Vv||^2 + \lambda(||v||^2 - \epsilon\, ||T^{\dagger}\xi||^2)$$

it is easily shown that

$$J(v + \delta v) - J(v) = 2 < V^*\delta\xi + V^*Vv + \lambda v, \delta v >.$$

Hence, at the optimal v, the equation

$$V^*\delta\xi + V^*Vv + \lambda v = 0$$

must hold. The optimal control must obviously satisfy

$$v = -(\lambda I + V^*V)^{-1} V^*\delta\xi, \ \lambda > 0 \tag{1}$$

where the scalar λ has yet to be determined.

From equation 1 it follows that

$$||v||^2 = < \delta\xi, \ V(\lambda I + V^*V)^{-2} \ V^*\delta\xi > = < \delta\xi, \ (\lambda I + VV^*)^{-2} VV^*\delta\xi >$$

and hence that λ must satisfy the condition

$$< \delta\xi, \ (\lambda I + VV^*)^{-2} VV^*\delta\xi > = \epsilon \ ||T^{\dagger} \xi||^2. \tag{2}$$

For $\delta\xi \in R(V)^{\perp}$ it is well known that $V^*\delta\xi = 0$ and hence equation 2 can be satisfied only if $\epsilon = 0$. This is the case where the control $v = 0$ is optimal for our problem. Moreover it is not difficult to show that $\overline{R(V)}$ and $R(V)$ are reducing subspaces for $(\lambda I + VV^*)^{-2}$, hence only the component of $\delta\xi$ in $\overline{R(V)}$ effects the right hand side of equation 2. Rather than taking projections it will be assumed that $\delta\xi \in \overline{R(V)}$.

2. Two Other Specific Problems. In the next two problems the functional J defined by

$$J(u) = ||u||^2 + \alpha \ ||\delta Tu||^2$$

plays an important role. Here the energy of a control and the system output disturbance corresponding to that control are weighted together to form the functional.

Problem 2. Assume for $\xi \in H_2$ that $||T^{\dagger} \xi|| < k$. Find the largest α for which a control u of norm k exists such that u minimizes J while satisfying $\xi = Tu$.

In other words we wish to determine the maximum weight that can be given to the output disturbance such that a control exists, which is optimal with respect to J and the constraint, while having acceptable energy. In analyzing this problem advantage will be taken of the fact that the bilinear functional

$$(x,y) = < x, \ (I + \alpha \ \delta T^*\delta T)y > \ x,y \in H_1$$

is an inner product on H_1. Moreover the norm induced by this inner product is precisely the functional J (this new norm is actually equivalent to the given norm). Thus, if $T^{\#}$ denotes the pseudo inverse of T with respect to the norm J, problem 2 may be rephrased as finding the largest α for which $||T^{\#}\xi|| = k$.

For convenience the positive self adjoint operator $K = \delta T^*\delta T$ is introduced. Then it can be shown (see [9, Chapter IV]) that

$$T^{\#} = (I + \alpha K)^{-1} \ T^* \ [T \ (I + \alpha K)^{-1} \ T^*]^{-1} \tag{3}$$

where T^* and δT^* still denote adjoints with respect to the original norm. (Since $T^{\circledast}$, the adjoint of T with respect to the new norm, satisfies $T^{\circledast} = (I + \alpha K)^{-1} \ T^*$ the familiar identification $T^{\#} = T^{\circledast}(TT^{\circledast})^{-1}$ holds as usual.)

In the following lemma, T is onto and P denotes the orthogonal projection of H onto N(T).

Lemma 1. Whenever the indicated inverse exists

$$T^{\#} = (I + \alpha PK)^{-1} \ T^{\dagger}$$

holds.

The proof of this lemma, which is quite straight forward, may be found in [12]. Before proceeding with the solution of Problem 2, consider the following related problem.

Problem 3. In Problem 2 set $k^2 = (1 + \epsilon)\ ||T^\dagger \xi||^2$ and determine α from the equality $J(T^\# \xi) = (1 + \epsilon)\ ||T^\dagger \xi||^2$. Problem 3 evidently reduces to a study of the scalar equation

$$||T^\# \xi||^2 + \alpha ||\delta T T^\# \xi||^2 = (1 + \epsilon)\ ||T^\dagger \xi||^2. \tag{5}$$

Recalling that $\delta T^* \delta T = K$ this equation becomes

$$< T^\# \xi,\ (I + \alpha K) T^\# \xi > = (1 + \epsilon) < T^\dagger \xi,\ T^\dagger \xi > .$$

Using lemma 1 we have

$$< T^\dagger \xi,\ [(I + \alpha KP)^{-1} (I + \alpha K)\ (I + \alpha PK)^{-1} - (1 + \epsilon) I]\ T^\dagger \xi > = 0$$

which may be rewritten as

$$< \xi,\ (T^\dagger)^* W(\alpha) T^\dagger \xi > - \epsilon < \xi,\ (T^\dagger)^* T^\dagger\ \xi > = 0$$

where

$$W(\alpha) = (I + \alpha KP)^{-1}\ (I + \alpha K)(I + \alpha PK)^{-1} - I. \tag{6}$$

Thus a corollary of lemma 1 is the following:

Corollary. The dependence between α and ϵ is determined by

$$< \xi,\ (T^\dagger)^*\ W(\alpha) T^\dagger \xi > = \epsilon < \xi,\ (T^\dagger)^* T^\dagger\ \xi >. \tag{7}$$

Lemma 2. Whenever $I + \alpha PK$ has a bounded inverse

(a) $I + \alpha KP$ has a bounded inverse

(b) $I - (I + \alpha KP)^{-1} = \alpha KP(I + \alpha KP)^{-1}$

(c) $(I + \alpha KP)^{-1}\ (I + \alpha PK)^{-1} - (I + \alpha PK)^{-1} - (I + \alpha KP)^{-1} + I = \alpha^2 KPK(I + \alpha PK)^{-2}$.

Proof. Part (a) follows easily from the identity

$$(I + \alpha KP)^{-1} = I - \alpha K(I + \alpha PK)^{-1} P$$

where first the right hand side defines the left and then it is shown that the left hand side is the inverse of $I + \alpha KP$. Similarily (b) follows by clearing fractions. To prove (c) first rewrite the left hand side of this identity to obtain

$$[I - (I + \alpha KP)^{-1}][I - (I + \alpha PK)^{-1}] = \alpha^2 KPK(I + \alpha PK)^{-2}$$

Then using (b) this becomes

$$\alpha KP(I + \alpha KP)^{-1}\ \alpha PK(I + \alpha PK)^{-1} = \alpha^2 KPK(I + \alpha PK)^{-2}$$

which is equivalent to

$$(I + \alpha KP)^{-1}\ KPK = KPK(I + \alpha PK)^{-1}$$

This last equality is easily verified completing the proof.

Corollary. $(T^\dagger)^* W(\alpha) T^\dagger = \alpha (T^\dagger)^* [K(I + \alpha PK)^{-1}] T^\dagger$

Proof. Using definition (equation 6) of $W(\alpha)$ and some elementary operations the equality chain

$$\begin{aligned} W(\alpha) &= (I + \alpha KP)^{-1}\ (I + \alpha K)(I + \alpha PK)^{-1} - I \\ &= (I + \alpha KP)^{-1}\ (I + \alpha PK)^{-1} - (I + \alpha KP)^{-1} - (I + \alpha PK)^{-1} + I \\ &\quad + (I + \alpha KP)^{-1} + (I + \alpha PK)^{-1} + \alpha\ K(I + \alpha PK)^{-2} - 2I \end{aligned}$$

is established. Using part (c) of lemma 2 one obtains

$$
\begin{aligned}
W(\alpha) &= \alpha^2 KPK\,(I + \alpha PK)^{-2} + \alpha K(I + \alpha PK)^{-2} + (I + \alpha KP)^{-1} + (I + \alpha PK)^{-1} - 2I \\
&= \alpha K\,(I + \alpha PK)(I + \alpha PK)^{-2} + (I + \alpha KP)^{-1} + (I + \alpha PK)^{-1} - 2I \\
&= \alpha K(I + \alpha PK)^{-1} + (I + \alpha KP)^{-1} + (I + \alpha PK)^{-1} - 2I.
\end{aligned}
$$

The corollary now follows from this equation and the fact that

$$(I + \alpha KP)^{-1}\, T^{\dagger} = T^{\dagger}.$$

To verify this latter assertion recall that P projects on $N(T)$ while the range of $T^{\dagger}$ is in $N(T)^{\perp}$. Therefore $PT^{\dagger} = 0$ and hence

$$T^{\dagger} = (I + \alpha KP)^{-1}\,(I + \alpha KP)T^{\dagger} = (I + \alpha KP)^{-1}\, T^{\dagger}.$$

Since K and P are self adjoint this equality implies also that

$$(T^{\dagger})^* = (T^{\dagger})^*\,(I + \alpha PK)^{-1}$$

and hence

$$(T^{\dagger})^*[(I + \alpha KP)^{-1} + (I + \alpha PK)^{-1} - 2I]\, T^{\dagger} = 0$$

which completes the proof of the corollary. An immediate consequence is that in Problem 3 the dependence of α on ϵ is expressed by the equation

$$\alpha < T^{\dagger}\xi,\ K(I + \alpha PK)^{-1}\, T^{\dagger}\xi > = \epsilon\ ||T^{\dagger}\xi||^2 \tag{8}$$

Consider now Problem 2 with $k^2 = (1 + \epsilon)\ ||T^{\dagger}\xi||$. Evidently it suffices to study the equation

$$< T^{\dagger}\xi,\ (I + \alpha KP)^{-1}\,(I + \alpha PK)^{-1}\, T^{\dagger}\xi > = (1 + \epsilon) < T^{\dagger}\xi,\ T^{\dagger}\xi >.$$

Since

$$(I + \alpha KP)^{-1}\,(I + \alpha PK)^{-1} - I = W(\alpha) - \alpha K(I + \alpha PK)^{-2}$$

it follows from the corollary to lemma 2 that

$$
\begin{aligned}
< T^{\dagger}\xi,\ [(I + \alpha KP)^{-1}\,(I + \alpha PK)^{-1} - I]\, T^{\dagger}\xi > &= < T^{\dagger}\xi,\ [\alpha K(I+\alpha PK)^{-1} \\
&\quad - \alpha K(I+\alpha PK)^{-2}]T^{\#}\xi > = \alpha^2 < T^{\dagger}\xi,\ KPK(I + \alpha PK)^{-2}\, T^{\dagger}\xi >.
\end{aligned}
$$

Thus in Problem 2 the dependence of α on ϵ is determined by

$$\alpha^2 < T^{\dagger}\xi,\ KPK(I + \alpha PK)^{-2}\, T^{\dagger}\xi > = \epsilon\ ||T^{\dagger}\xi||^2. \tag{9}$$

3. Comparing the Solutions. In comparing the solutions of Problems 1, 2 and 3 it is helpful to rework the results for Problem 1. Recall first that $\delta\xi = \delta Tu_{\xi} = \delta TT^{\dagger}\xi$ and that $V = \delta TP$. Thus it follows that

$$V^*V = P\delta T^*\delta TP = PKP$$

and that equation 1 may be rewritten as

$$v = -(\lambda I + PKP)^{-1}\,(P\delta T^*)(\delta TT^{\dagger}\xi) = -PK(\lambda I + PK)^{-1}\, T^{\dagger}\xi.$$

Let $T^{\ddagger}$ denote the mapping of ξ (for every ξ) to the solution of Problem 1. Then since $T^{\ddagger}\xi = T^{\dagger}\xi + v$ it follows that

$$T^{\ddagger} = [I - PK(\lambda I + PK)^{-1}]T^{\dagger}.$$

If $\lambda = 1/\beta$ observe that $(\lambda I + PK)^{-1} = \beta(I + \beta PK)^{-1}$ and that the identity

$$(I + \beta PK)^{-1} = I - \beta PK(I + \beta PK)^{-1}$$

holds. Thus $T^{\ddagger}$ may be identified as in the equality

$$T^{\ddagger} = (I + \beta PK)^{-1}\, T^{\dagger}.$$

Consider now equation 2. By direct substitution and some elementary rearrangements the equality

$$< \delta\xi, V(I + V^*V)^{-2}V^*\delta\xi > = \beta^2 < T^\dagger\xi, KPK(I + \beta PK)^{-2} T^\dagger\xi >$$

holds. It therefore follows that β must satisfy the equality

$$\beta^2 < T^\dagger\xi, KPK(I + \beta PK)^{-2} T^\dagger\xi > = \epsilon \, ||T^\dagger\xi||^2. \tag{11}$$

In view of equations 3 and 10 it is obvious that both $T^\#$ and $T^\ddagger$ have the form $(I + \gamma PK)^{-1} T^\dagger$. In view of equations 9 and 11 it is obvious that the solutions of Problems 1 and 2 are identical. The form of the optimal control in Problem 3 is the same as in Problems 1 and 2. The value of γ, however, is different being determined by equation 8.

It is also of interest to compute the norm of E_ξ, the actual miss distance. Now recalling that P has range in N(T) it follows that

$$T(I + \gamma PK)^{-1} = T$$

holds for all γ. This, together with the identity $TT^\dagger\xi = \xi$, establishes the equality chain

$$E_\xi = \xi - (T + \delta T)(I + \gamma PK)^{-1} T^\dagger\xi = \delta T(I + \gamma PK)^{-1} T^\dagger\xi.$$

Hence it follows that

$$||E_\xi||^2 = ||\delta T(I + \gamma PK)^{-1} T^\dagger\xi||^2$$

Thus, in all three problems the system errors have the same form differing only in the value of the scalar γ.

Since the distinction between the three problems hinges on the two scalars involved a comparison of equations 8 and 11 is of interest. Since $K = \delta T^*\delta T$ equations 8 and 11 may be rewritten as

$$f_2(\alpha) = \alpha < \eta, (I + \alpha Q)^{-1}\eta > = \epsilon \, ||T^\dagger\xi||^2$$

$$f_1(\beta) = \beta^2 < \eta, Q(I + \beta Q)^{-2}\eta > = \epsilon \, ||T^\dagger\xi||^2$$

respectively where $\eta = \delta\xi = \delta TT^\dagger\xi$ and $Q = \delta TP\delta T^*$. In [12] it is shown that both f_1 and f_2 are monotone increasing (or monotone nondecreasing) functions on $[0,\infty)$. Furthermore it is shown that $f_1(\gamma) \geq f_2(\gamma)$ holds for $\gamma \in [0,\infty)$. Consequently if α, β are solutions to the equations

$$f_2(\alpha) = \epsilon \, ||T^\dagger\xi||^2 = f_1(\beta)$$

for any $\epsilon \geq 0$ then $\alpha \leq \beta$. Now in [12] it is also shown that

$$||E_\xi||^2 = < \eta, (I + \gamma Q)^{-2}\eta > = f_4(\alpha)$$

is a monotone decreasing function with respect to γ. Therefore the solution to problems 1 and 2 is always better than or at least equal to the solution of Problem 3 as far as system errors are concerned.

Example 1. In this example two degenerate cases are examined. First consider the case where the system disturbance takes the form of a pure gain change. That is $\delta T = \mu T$ for some scalar $\mu \neq 0$. Then $K = \mu^2 T^*T$ and since P annihilates T^* it follows that $PK = o$ and hence

$$T^\ddagger = T^\# = T^\dagger.$$

This result verifies the obvious fact that the constraint $Tu = \xi$ fixes $\delta\xi = \delta Tu = \mu Tu = \mu\xi$ and hence no minimizing of sensitivity is possible. More generally if δT is a distortion (see [6] or [10]) then $PK = 0$ and the same conclusion holds.

As a second specific case assume that P and K commute. Then

$$T^\# = T^\ddagger = (I + \gamma \, PK)^{-1}T^\dagger = (I + \gamma \, KP)^{-1}T^\dagger = T^\dagger$$

holds for any γ. Also

$$||E_\xi||^2 = ||\delta T(I + \gamma\ PK)^{-1}T^\dagger \xi||^2 = ||\delta T\ T^\dagger \xi|| = ||\delta\xi||^2$$

holds for all γ and once more it is impossible to reduce the nominal disturbance. If $N(T)^\perp \subset N(\delta T)$ then $KP = PK$. To verify this assertion note that $N(T)^\perp \subset N(\delta T)$ implies $R(I - P) \subset N(\delta T)$ which implies $K(I - P) = 0$. Taking adjoints shows that $(I - P)K = 0$ and hence subtracting we have

$$0 = K(I - P) - (I - P)K = PK - KP$$

which verifies the assertion. It was noted earlier that when $N(T)^\perp \subset N(\delta T)$ then $\delta\xi = \delta T T^\dagger \xi = 0$ which can't be improved on.

Example 2. Another case which is relatively easy to analyze is that where T and δT are functionals. Then $P = I - ||T||^{-2}\ T^*T$ and

$$Q = \delta T\ P\delta T^* = \delta T\delta T^* - ||T||^{-2}\ \delta TT^*T\ \delta T^*$$
$$= ||\delta T||^2 - ||T||^{-2}\ |< T,\ \delta T >|^2 = ||\delta T||^2 \sin^2\theta$$

where θ is the angle between T and δT (i.e. $\cos\theta = < T,\ \delta T > / ||T||\cdot||\delta T||$).

Consequently,

$$f_2(\alpha) = \alpha(1 + \alpha\ ||\delta T||^2 \sin^2\theta)^{-1}\ ||\delta\xi||^2$$
$$f_1(\beta) = \beta^2(1 + \beta\ ||\delta T||^2 \sin^2\theta)^{-2}\ ||\delta\xi||^2$$
$$||E_\xi||^2 = (1 + \gamma\ ||\delta T||^2 \sin^2\theta)^{-2}\ ||\delta\xi||^2.$$

Using these equations $||E_\xi||^2$ may be explicitely determined as a function of ϵ. For instance for problem 3 the equation $f_2(\alpha) = \epsilon||T^\dagger\xi||^2$ provides the relationships

$$\alpha = (||\delta\xi||^2 - \epsilon\ ||T^\dagger\xi||^2.\ ||\delta T||^2 \sin^2\theta)^{-1}\ \epsilon\ ||T^\dagger\xi||^2$$
$$(1 + \alpha\ ||\delta T||^2 \sin^2\theta) = \alpha\ ||\delta\xi||^2/\ \epsilon\ ||T^\dagger\xi||^2$$

from which it follows that

$$(1 + \alpha\ ||\delta T||^2 \sin^2\theta) = ||\delta\xi||^2\ (||\delta\xi||^2 - \epsilon\ ||T^\dagger\xi||^2\cdot\ ||\delta T||^2 \sin^2\theta)^{-1}$$

and consequently

$$||E_\xi|| = (||\delta\xi||^2 - \epsilon\ ||T^\dagger\xi||^2\cdot\ ||\delta T||^2 \sin^2\theta)/\ ||\delta\xi||.$$

This equation, or the inequality

$$||E_\xi||/\ ||\delta\xi|| \leq 1 - \epsilon \sin^2\theta$$

(which follows easily from it) establishes a simple relationship between controller surplus capacity and the potential percentage reduction in disturbance.

Example 3. Consider the linear dynamic system satisfying the vectoral differential equation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(t) = x^o.$$

This system is to perform a fixed time state transfer $(t_o, x^o) \to (t_f, x^f)$ with minimum energy. The transformation T is identified by the equation

$$Tu = \int_{t_o}^{t_f} \Phi(t_f, s)B(s)u(s)ds$$

while $\xi = x^f - \Phi(t_f, t_o)x^o$. Similarly T^* and $T^\dagger$ can be directly computed. The matrix B is assumed to depend on a parameter μ and that by using a linearization this dependence can be expressed in the form

$$B(t,u) = B_1(t) + \mu B_2(t)$$

where B_1 and B_2 are fixed (known) matrices. T and δT are both defined as above with $B = B_1$ and $B = B_2$ in the two respective cases.

For any control u,

$$||\delta\xi||^2 = \mu^2 \; || \int_{t_o}^{t_f} \Phi(t_f,s)B_2(s)u(s)ds||^2.$$

holds. When μ is a random variable with zero mean $\delta\xi$ is a random vector with zero mean. The standard deviation σ_μ of μ and Σ_ξ^2 the expectation of $||\delta\xi||^2$, are related by

$$\Sigma_\xi^2 = \sigma_\mu^2 \; || \int_{t_o}^{t_f} \Phi(t_f,s)B_2(s)u(s)ds||^2.$$

If the controller has surplus capacity it is natural to attempt a partial minimization of Σ_ξ. It is obvious that such a problem falls within the framework of the present analysis. To complete the example the necessary computations will be outlined.

If u and x are m and n tuples respectively the Hilbert spaces $H_1 = [L_2(t_o,t_f)]^m$ and $H_2 = E^n$ are the natural setting for the problem. It can be easily shown (see [9, sec 3.3]) that T* is computed by the equation

$$(T^*\lambda)(t) = B_1^*(t)\Phi^*(t_f,t)\lambda, \quad t \in [t_o,t_f], \; \lambda \in E^n$$

while δT^* has the same form with B_2 replacing B_1. Since $P = I - T^\dagger T$ and assuming that the scalar γ is small it follows that

$$(I + \gamma PK)^{-1}T^\dagger = (I - \gamma PK)T^\dagger = T^\dagger - \gamma \; KT^\dagger + \gamma T^\dagger TKT^\dagger .$$

Letting N and M denote the matrices

$$N = T\delta T^* = \int_{t_o}^{t_f} \Phi(t_f,s)B_1(s)B_2^*(s)\Phi^*(t_f,s)ds$$

$$M = TT^* = \int_{t_o}^{t_f} \Phi(t_f,s)B_1(s)B_1^*(s)\Phi^*(t_f,s)ds$$

respectively (where M is assumed to be nonsingular) the components of $(I + \gamma \; PK)^{-1}T^\dagger$ may be identified as

$$(T^\dagger \xi)(t) = B_1^*(t) \; \Phi^*(t_f,t)M^{-1}\xi, \qquad t \in [t_o,t_f]$$

$$(KT^\dagger \xi)(t) = B_2^*(t) \; \Phi^*(t_f,t)N^*M^{-1}\xi, \qquad t \in [t_o,t_f]$$

$$(T^\dagger TKT^\dagger)(t) = B_1^*(t) \; \Phi^*(t_f,t)M^{-1}NN^*M^{-1}\xi, \qquad t \in [t_o,t_f]$$

The specification of the open loop controller is complete once γ is determined. The only distinction here with the usual case is that $\sigma_\mu^2\alpha$ (respectively $\sigma_\mu^2\beta$) must be used in place of α (respectively β) in the determining equations.

Remark. When x and u are scalar valued functions examples 2 and 3 may be combined (similar results hold for all single input single output systems). Using the results of example 2 it follows that

$$\frac{\Sigma_\xi}{||\delta\xi||\sigma_\mu} \leq (1 - \epsilon \sin^2 \theta)$$

reflects the statistical improvement in the system performance.

4. Discussion. In Section 3 the comparison of Problems 1 and 2 with Problem 3 is apriori biased in favor of Problems 1 and 2. Indeed, in the first two cases the energy constraint is of the form

$$||T^{\#}\xi||^2 = (1 + \epsilon)\ ||T^{\dagger}\xi||^2$$

while in the latter case

$$||T^{\#}\xi||^2 + \alpha||\delta TT^{\#}\xi||^2 = (1 + \epsilon)\ ||T^{\dagger}\xi||^2$$

must hold. Clearly the second constraint is more restrictive then the first and hence the first case should be expected to result in a smaller system error. Thus Section 3 should be viewed as a comparison of the properties of the different constraints and not as a comparison of alternatives to the same problem.

Except in Example 2 when T and δT are functionals, the functional dependence of α on ϵ has been left in implicit form. Due to the monotonic behavior and the differentiability of the functions f_1 and f_2 it appears that the computational problem involved is only mildly difficult. For small ϵ, α, β and γ the first order approximations

$$f_2(\alpha) = \alpha\ ||\delta TT^{\dagger}\xi||^2 = \epsilon\ ||T^{\dagger}\xi||^2$$

$$f_1(\beta) = \beta^2\ ||PKT^{\dagger}\xi||^2 = \epsilon\ ||T^{\dagger}\xi||^2$$

$$||E_{\xi}||^2 = ||\delta TT^{\dagger}\xi||^2 - 2\gamma||PKT^{\dagger}\xi||^2$$

hold. Combining these equations results in

$$||E_{\xi}||^2 = ||\delta\xi||^2 - 2\ \sqrt{\epsilon}\ ||T^{\dagger}\xi||\cdot||PKT^{\dagger}\xi||$$

for Problems 1 and 2 and

$$||E_{\xi}||^2 = ||\delta\xi||^2 - 2\epsilon\ ||T^{\dagger}\xi||^2$$

for Problem 3. These two system error equations show an obvious difference in the dependence between $||E_{\xi}||^2$ and ϵ.

REFERENCES

1. Pagurek, B., "Sensitivity of the Performance of Optimal Control Systems to Plant Parameter Variations," IEEE Trans. on Automatic Control, Vol. AC 10, pp. 178-180, April 1965.
2. Witsenhausen, H. S., "On the Sensitivity of Optimal Control Systems," IEEE Trans. on Automatic Control (Correspondence) Vol. AC-10, No. 4, pp. 495-496, October 1965.
3. Holtzman, J. M., Horing, S., "The Sensitivity of Terminal Conditions of Optimal Control Systems to Parameter Variations," IEEE Trans. on Automatic Control, Vol. AC-10, No. 4, October 1965.
4. Porter, W. A., "On the Sensitivity of Terminal Conditions," IEEE Trans. on Automatic Control, Vol. AC-11, No. 3, July 1966.
5. Rohrer, R. A., Sobral, M., "Sensitivity Considerations in Optimal System Design," IEEE Trans. on Automatic Control, Vol. AC-10, No. 1, pp. 43-48, January 1965.
6. Porter, W. A., "Sensitivity Problems in Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-10, No. 3, pp. 301-307, July 1965.
7. Perkins, W. R., and J. B. Cruz, Jr., "The Parameter Variation Problem in State Feedback Control Systems," presented at the 1964 Joint Automatic Control Conference, Stanford University, Stanford, California.
8. Porter, W. A., "Some Theoretical Limitations of System Sensitivity Reduction," Proceedings Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, October 1965.
9. Porter, W. A., Modern Foundations of System Analysis, Macmillan Co., New York, 1966.
10. Porter, W. A., "Concerning Disturbance Measures in Linear Systems" IEEE Trans. on Automatic Control, Vol. AC-11, No. 3, July 1966.
11. Taylor, A. E., Introduction to Functional Analysis, John Wiley, New York, 1961.
12. Porter, W. A., "Improving System Sensitivity via Surplus Controller Capacity," Journal of the Franklin Institute, December, 1966.

SYNTHESIS FOR MINIMUM SENSITIVITY UNDER WORST CASE CONDITIONS*

RICHARD LAWRENCE GONZALES
Bradley University, Peoria, Illinois

ABSTRACT

A new multiparameter sensitivity index for control systems is introduced. This measure of the deterioration in the system performance is based either on the integral square error or the mean square error in the system response due to parameter variations. A first order approximation to the change in the system transmission is used to simplify these functions.

A synthesis procedure is derived which permits the minimization of the control system sensitivity with respect to the plant parameters under worst case conditions. This procedure is illustrated by a design example and the errors due to the first order approximation are analyzed.

INTRODUCTION

Feedback is utilized in control systems primarily for two reasons:

1) It may decrease the effects of parameter variations on the system behavior.
2) It may reduce unwanted effects due to noise or disturbance signals.

This paper is concerned with the first of these reasons.

The sensitivity problem in control systems arises because of a lack of knowledge of the values of some of the systems parameters. In general, however, the range of these parameters and their nominal values may be estimated or determined. The specific problem to be considered here is the development of analysis and synthesis procedures for linear time invariant control systems to minimize the effects on the system behavior when several parameters vary simultaneously. The problems associated with multiparameter sensitivity have been considered by a number of researchers. Significant contributions have been made by Goldstein and Kuo [3], Hakimi and Cruz [5], Remec [10], Mikulski [8], and Blostein [1]. It is convenient here to treat the multiparameter case in terms of a logical extension of the Bode sensitivity function for single parameter systems [3,8]. This extension results in an expression which is a first order approximation to the change in the control system transfer function due to parameter variation. Mazer [7] and Rutman [11] have proposed an integrated square or mean square measure of the effects of parameter variation. Mazer [7] utilizes Bode's original sensitivity function and thus only considers the single parameter system. Rutman [11] has proposed that Mazer's technique could be entended to multiparameter systems. However, he has not developed procedures for the implementation of these ideas.

---

*This paper is based on a portion of a dissertation submitted in partial fulfillment of the requirements for the doctoral degree at the University of Illinois. The work described was done at the Coordinated Science Laboratory of the University of Illinois under contract DA 28 043 AMC 0073 (E).

The following is a brief outline of this paper. The concepts of a multiparameter sensitivity index are introduced. These concepts lead to a system sensitivity index which is fundamental to the developments which follow. Synthesis techniques for control systems under worst case conditions are developed which permit the minimization of this system sensitivity index. These techniques are applicable when information concerning the plant parameter deviations are expressed by tolerances. Finally, a design example is discussed.

THE SYSTEM SENSITIVITY INDEX

A new measure of multiparameter sensitivity will be developed in this section which will be referred to as the system sensitivity index. This index is based on the assumption that the integral square or mean square value of the error in the system response due to plant parameter variation is a reasonable measure of sensitivity. The index incorporates a first order approximation of the Bode sensitivity function to mathematically express the change in the system transfer function due to plant parameter perturbations. Thus, the system sensitivity index becomes a function of the plant parameter deviations and the unconstrained numerator coefficients of the Bode sensitivity function. This final result is essentially an extension of Mazer's work [7].

The control system model which will be considered is shown in Figure 1.

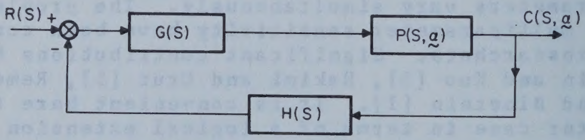


Figure 1. Control System Model

The compensating network G(s) and H(s) are considered to be parameter invariant. The only component subject to parameter variations is thus $P(s,\underset{\sim}{\alpha})$. The parameters $\underset{\sim}{\alpha}$ may be expressed by

$$\underset{\sim}{\alpha} = \underset{\sim}{\alpha}_o + \Delta\underset{\sim}{\alpha} . \qquad (1)$$

where $\underset{\sim}{\alpha}_o$ is the nominal or design value and $\Delta\underset{\sim}{\alpha}$ is the deviation from this nominal value. From Figure 1 the overall system transfer function $T(s,\underset{\sim}{\alpha})$ may be obtained.

$$T(s,\underset{\sim}{\alpha}) = \frac{P(s,\underset{\sim}{\alpha})G(s)}{1 + P(s,\underset{\sim}{\alpha})G(s)H(s)} = \frac{N(s,\underset{\sim}{\alpha})}{D(s,\underset{\sim}{\alpha})} \tag{2}$$

The Bode sensitivity function [2] for the control system of Figure 1 may be written as

$$S^T_P \triangleq S^T_P(s,\underset{\sim}{\alpha},\underset{\sim}{B}) = \frac{1}{1 + P(s,\underset{\sim}{\alpha})G(s)H(s)} = \frac{\sum\limits_{k=0}^{M} B_k s^k}{D(s,\underset{\sim}{\alpha})} = \frac{dT}{dP}\,\frac{P}{T} \tag{3}$$

It is to be noted that the numerator and denominator of (3) must be polynomial in s which are of the same order M. The compensator G(s) and H(s) are uniquely specified by $T(s,\underset{\sim}{\alpha}_o)$, $S^T_P(s,\underset{\sim}{\alpha}_o)$ and $P(s,\underset{\sim}{\alpha}_o)$. That is,

$$G(s) = \frac{T(s,\underset{\sim}{\alpha}_o)}{S^T_P(s,\underset{\sim}{\alpha}_o,\underset{\sim}{B})P(s,\underset{\sim}{\alpha}_o)} = \frac{N(s,\underset{\sim}{\alpha}_o)}{\left[\sum\limits_{k=0}^{M} B_k s^k\right] \times P(s,\underset{\sim}{\alpha}_o)}, \tag{4}$$

and

$$H(s) = \frac{1 - S^T_P(s,\underset{\sim}{\alpha}_o,\underset{\sim}{B})}{T(s,\underset{\sim}{\alpha}_o)} = \frac{D(s,\underset{\sim}{\alpha}_o) - \sum\limits_{k=0}^{M} B_k s^k}{N(s,\underset{\sim}{\alpha}_o)} \tag{5}$$

For a desired nominal transfer function $T(s,\underset{\sim}{\alpha}_o)$ one may thus select $\underset{\sim}{B}$ such that S^T_P minimizes the desired system sensitivity index. Constraints are usually imposed on $\underset{\sim}{B}$ such that the system is stable for all perturbations of $\underset{\sim}{\alpha}$ in the tolerance band and so that the compensating network $\tilde{G}$ and H are of the desired form.

The system sensitivity index will now be defined either as the integral of the square of response error due to plant variations or as the mean square value of this response error.

$$I \triangleq \int_{\infty}^{\infty} [e(t,\Delta\underset{\sim}{\alpha},\underset{\sim}{B})]^2 dt, \text{ or } \lim_{\tau\to\infty} \frac{1}{2\tau} \int_{\tau}^{\tau} [e(t,\Delta\underset{\sim}{\alpha},\underset{\sim}{B})]^2 dt, \tag{6}$$

where I is the system sensitivity index and e is the response error due to plant parameter deviations from their nominal values.

Two measures of sensitivity have been defined and are expressed by (6); both of these expressions will be referred to as the system sensitivity index. The major advantages of these definitions are:

1. They yield a measure of sensitivity which is independent of frequency.
2. They lend themselves to a synthesis procedure which will be developed.
3. They provide a means of comparing different control system designs. The smaller $I(\Delta\alpha,\underset{\sim}{B})$ the less sensitivity the system. The minimum possible value of $I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})$ is zero and this corresponds to no sensitivity to parametric variation.

Some difficulties associated with these definitions of the system sensitivity index are:

1. They may not satisfy a specific performance specification to which a designer is required to conform. For example, a specific problem may require that emphasis be placed on the static accuracy.
2. The weighting of large errors more heavily than small errors may not be desirable in some applications.
3. These measures of sensitivity are applicable only when deterministic or stochastic properties of the control system input are known. In other words, the system sensitivity index may not be obtained for the system alone. This disadvantage may be overcome in some problems by introducing a ficticious input such as an impulse.

A first order approximation to the change in the system transfer function may be obtained from (3). This approximation is expressed by

$$\Delta T(s,\Delta\underset{\sim}{\alpha},\underset{\sim}{B}) = S^{T}_{P}\Bigg|_{\alpha = \alpha_o} \frac{T(s,\underset{\sim}{\alpha}_o)}{P(s,\underset{\sim}{\alpha}_o)} \sum_{i=1}^{N} \frac{\delta P(s,\underset{\sim}{\alpha}_o)}{\delta\alpha_i}\Delta\alpha_i \qquad (7)$$

Thus a first order approximation to the system sensitivity index may be found using (3), (6), (7), and Parsevals Theorem [9].

$$I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B}) \cong \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty} F(s,\underset{\sim}{\alpha}_o)F(-s,\underset{\sim}{\alpha}_o)\Bigg[\sum_{k=0}^{M} B_k s^k \sum_{\ell=0}^{M} B_k(-s)^k$$

$$\sum_{i=1}^{N}\frac{\delta P(s,\underset{\sim}{\alpha}_o)}{\delta\alpha_i}\Delta\alpha_i \sum_{j=1}^{N}\frac{\delta P(-s,\underset{\sim}{\alpha}_o)}{\delta\alpha_j}\Delta d_j\Bigg]\, ds, \qquad (8)$$

where

$$F(s,\underset{\sim}{\alpha}_o) \triangleq \frac{T(s,\underset{\sim}{\alpha}_o)}{P(s,\underset{\sim}{\alpha}_o)D(s,\underset{\sim}{\alpha}_o)}\left\{\begin{matrix} R(s) \\ \text{or} \\ \Phi^{+}_{RR}(s)\end{matrix}\right\}, \qquad (9)$$

and the power spectral density of the input signal is given by

$$\Phi_{RR}(s) \overset{\Delta}{=} \Phi^{+}_{RR}(s)\ \Phi^{+}_{RR}(-s) \tag{10}$$

The analysis of control systems for a given perturbation $\Delta\underset{\sim}{\alpha}$, using this first order approximation, follows directly from these equations. The integrals which must be evaluated in this analysis have been tabulated by Newton et al [9].

SYNTHESIS UNDER WORST CASE CONDITIONS

In this section the analysis and synthesis of control systems under worst case conditions will be discussed. The design criterion will be the system sensitivity index defined by (8), (9), and (10). Worst case analysis and synthesis procedures are applicable when the information concerning the deviations in the plant parameters is in the form of tolerances. That is, tolerances on the parameters which may be expressed by

$$\Delta\underset{\sim}{\alpha}\ \varepsilon\ \overline{A}. \tag{11}$$

Worse case conditions are defined as the maximum value of $I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})$ with respect to $\Delta\underset{\sim}{\alpha}$ for $\Delta\underset{\sim}{\alpha}\ \varepsilon\ \overline{A}$, or

$$I_{max} \overset{\Delta}{=} \max_{\Delta\underset{\sim}{\alpha}\ \varepsilon\ \overline{A}} [I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})]. \tag{12}$$

When there are N plant parameters which are bounded by tolerances such that $\Delta\underset{\sim}{\alpha}\ \varepsilon\ \overline{A}$, it can be shown [4] that the maximum value of the first order approximation to system sensitivity index (equation (8)) with respect to the N plant parameters, can occur only at one of the 2^N combinations of $\Delta\underset{\sim}{\alpha}$ constructed from the extreme values of the $\Delta\underset{\sim}{\alpha}_i$'s. Thus, for any particular $\underset{\sim}{B}$ the worst case value of I may be found by calculating the value of I for the 2^N $\Delta\underset{\sim}{\alpha}$'s and choosing the largest one from this set. The analysis of any given control system under worst case conditions may be accomplished using this procedure.

Synthesis techniques for minimum sensitivity will now be discussed. The design objective is to minimize the worst case value of the system sensitivity index with respect to $\underset{\sim}{B}$. Expressing this mathematically

$$I_W \overset{\Delta}{=} \min_{\underset{\sim}{B}\ \varepsilon\ \beta} \left\{ \max_{\Delta\underset{\sim}{\alpha}\ \varepsilon\ \overline{A}} [I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})] \right\}. \tag{13}$$

It can be shown [4] that the term inside the curly braces in (13) has a response surface which is concave downward with respect to $\underset{\sim}{B}$. The value $\underset{\sim}{B}$ which minimizes the worst case value of the system sensitivity index is thus unique and may be determined by a number of optimum seeking methods. The pattern search method developed by Hooke and Jeeves [6] has been successfully used by the author for this purpose. Another method involving the solution of simultaneous algebraic equations may also be used in some cases [4].

A synthesis procedure has thus been developed which permits the minimization of the worst case value of the first order approximation to the integral square or mean square value of the error between the nominal system response and the perturbed system response. The error incurred due to the first order approximation will be analyzed for a particular design example which will be discussed next.

DESIGN EXAMPLE

In this section a specific problem will be used to illustrate some of the developments which have been discussed. A control system which has been designed by classical methods without any sensitivity considerations will be analyzed using the system sensitivity index as a measure of sensitivity. Another control system will be designed which yields the same transfer function as the classical design, however, in this design, the min-max synthesis procedures for a minimum system sensitivity index under worst case conditions will be applied. The classical, open loop, min-max, and nominal system designs will be compared with regard to sensitivity. Finally, an analysis of the errors involved in the first order approximation used for ΔT will be carried out for this example problem.

The control system is a data repeater system. The plant or fixed portion of the system consists of a motor and load combination which have the following specifications.

Motor

1. No load speed 5000 rpm minimum, 5600 rpm nominal, and 5800 rmp maximum.
2. J_m (inertia) = 1.3 gm - cm^2 ± 10%.
3. Stall torque = 0.6 oz-in ± 10%.
4. K_N (viscous friction coefficient) = 72 $\frac{\text{dyne-cm-sec}}{\text{radian}}$
5. Rated voltage 110 volts, 400 cycles.

Load

1. J_L (load inertia) = 13,000 gm-cm^2 ± 5%.
2. f_L (viscous friction) = 0.01 dyne-cm/rpm.
3. Gear ratio motor to load 1:30.

The motor was selected on the basis of its dynamic characteristics and the power requirements of the load.

The input signal will be an impulse for sensitivity consideration. A classical design which satisfies the desired or nominal performance specifications is obtained using a cascade lead network. Figure 2 shows a block diagram of this classical control system.

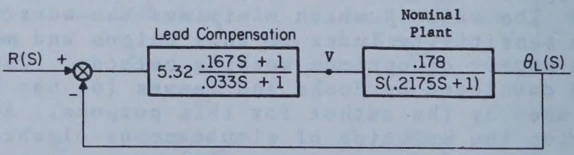


Figure 2. Classical Control System Design

From the specifications given for the motor and load combination, the linearized plant transfer function has the following form.

$$P(s,\underset{\sim}{\alpha}) = .000303 \quad \frac{\alpha_1\alpha_3}{S(\alpha_2\alpha_3 s + \alpha_1)} \tag{14}$$

where the nominal plant parameters and tolerances are respectively

$$\alpha_{1_o} = 42.6\text{x}10^3 \ , \ -4.26\text{x}10^3 \leq \Delta\alpha_1 \leq 4.26\text{x}10^3 \ , \tag{15}$$

$$\alpha_{2_o} = 15.74 \ , \ -.722 \leq \Delta\alpha_2 \leq .722 \ , \tag{16}$$

$$\alpha_{3_o} = 586 \ , \ -63 \leq \Delta\alpha_3 \leq 21. \tag{17}$$

Since $\underset{\sim}{B}$ is specified for the classical design, only the evaluation of the system sensitivity index at $2^N = 8$ points is required for worst case analysis. Evaluation of I at these points yields the following worst case value for the system sensitivity index

$$I_{wc} = .3342\text{x}10^{-2} \tag{18}$$

This result was obtained using a CDC 1604 digital computer. The same nominal transfer function as the classical design may also be obtained by an open loop structure. This open loop design was also analyzed and the resulting worst case value of the system sensitivity index is

$$I_{wo} = .5888\text{x}10^{-2} \tag{19}$$

Comparing the magnitudes of (18) and (19) indicates that the closed loop design is better than an open loop design.

For the plant specified it is now the objective to synthesize compensating networks G(s) and H(s) which yield the transfer function specified by the nominal classical design, and such that the system sensitivity index under worst case conditions is minimized. First, however, it is necessary to specify the B's in the numerator of the Bode sensitivity function that are to be constrained. From (4) and (5) one may write the general form of the compensating networks.

$$G(s) = \frac{s(21.75s + 130.3)(s + 4.6)}{.82(s^3 + B_2s^2 + B_1s + B_o)} \tag{20}$$

$$H(s) = \frac{(34.6-B_2)s^2 + (159.75 - B_1)s + (130.3 - B_o)}{21.75s + 130.3} \tag{21}$$

In order to avoid the problems associated with a differentiating element in the cascade compensator G(s), the variable B_o will be specified as zero.

The techniques previously discussed may now be used to determine the values of B_1 and B_2 such that the worst case value of the system sensitivity index is minimized. Again a computer solution was used and the resulting value of $\min\limits_{\underset{\sim}{B} \in \beta} [\max\limits_{\underset{\sim}{\Delta\alpha} \in \bar{A}} I(\underset{\sim}{\Delta\alpha},\underset{\sim}{B})]$ is,

$$I_{wm} = .6969 \times 10^{-4} \tag{22}$$

and

$$\underset{\sim}{B} = \begin{bmatrix} 0 \\ 7.61 \\ 0 \\ 1.0 \end{bmatrix} \tag{23}$$

A comparison of the numerical values of the system sensitivity index under worst case conditions shows that the open loop system is most sensitive, and the min-max design synthesized for minimum sensitivity, is by far the least sensitive. The classical design has a sensitivity which is better than the open loop design but worse than the min-max design.

As a method for comparing these designs further, the transient responses with an impulse input will now be considered for the nominal, open loop, classical and min-max designs. These time functions are plotted in Figure 3 for the worst case conditions. It is to be noted that these are true response curves for the plant perturbations and do not involve the first order approximation to ΔT which was used to calculate the system sensitivity index. An examination of these curves shows that the system sensitivity index has provided a good measure of the response error due to plant parameter variation. They also show that the min-max design procedure does yield a control system which is significantly less sensitive than the other designs.

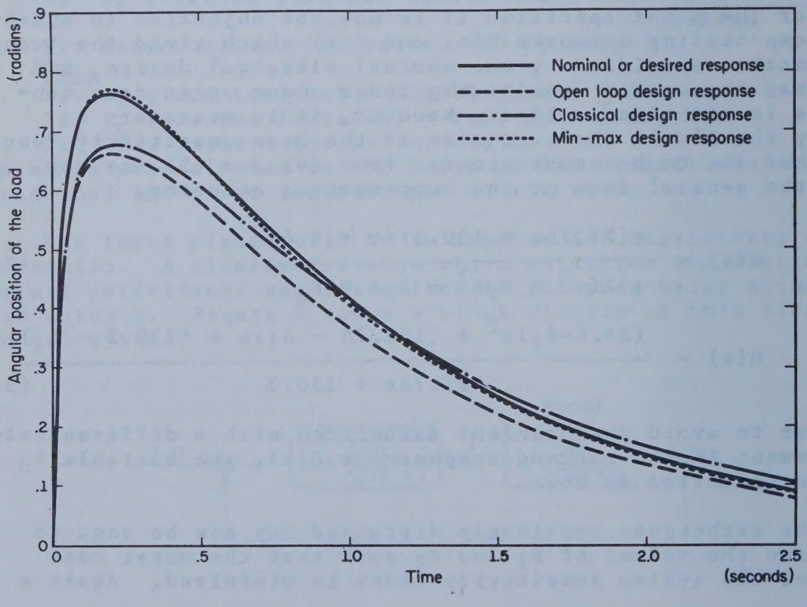


Figure 3. Impulse Response Under Worst Case Conditions

The analysis and synthesis carried out thus far is based on the first order approximation to the system sensitivity index given by (8). An exception to this statement is the transient response curves. It is now the objective to analyze the errors involved in using this first order approximation for the system sensitivity index. The approximation occurs due to first order approximations in ΔT given by (7). In order to calculate the true value of the system sensitivity index the actual value of ΔT may be found from the difference between the nominal and perturbed system transmission functions. Calculations of this type are also subject to error for small plant perturbations, because these calculations involve finding the difference of numbers of about the same amplitude. However, if perturbations from the nominal values are large or close to the worst case condition, the errors involved will be kept to a minimum. The discussion which follows leads to a direct computer search technique for determining the true value of $\min_{\underset{\sim}{B} \varepsilon \beta} [\max_{\Delta\underset{\sim}{\alpha} \varepsilon \underset{\sim}{A}} I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})]$, that is, one not involving a first order approximation.

The shape of the actual response surface of $I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})$ is unknown. However, it is reasonable to assume that the true min-max point is not very far from the one obtained using the first order approximation for $I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})$. It will be further assumed that in a small region about this point the true $I(\Delta\underset{\sim}{\alpha},\underset{\sim}{B})$ is unimodal in both $\Delta\underset{\sim}{\alpha}$ and $\underset{\sim}{B}$. One may thus use the approximate min-max point as a starting point and perform a direct search to determine the true min-max value of the system sensitivity index, and the corresponding values for $\Delta\underset{\sim}{\alpha}$ and $\underset{\sim}{B}$.

The pattern search technique of Hooke and Jeeves [6] must be modified to accomplish this objective. This is due to the fact that a min-max is required rather than a maximum or minimum. The modification involves perturbations from a given test point but excludes extrapolations. This exclusion does not cause any difficulties because the starting point is close to the solution and large extrapolations are thus unnecessary. This method has been programmed on the CDC 1604 computer and the true solution obtained for $\Delta\underset{\sim}{\alpha}$ and $\underset{\sim}{B}$ which minimizes the worst case conditions yields a difference only in value for B_1 (7.54 compared with 7.61). The difference in the numerical values of the system sensitivity index are larger ($.8221 \times 10^{-4}$ compared with $.6969 \times 10^{-4}$); however, these values do not have an absolute significance because the parameter $\underset{\sim}{B}$ is the only one which effects the compensating networks to be synthesized.

This example design problem thus demonstrates the effectiveness of the synthesis procedure. It also illustrates that in this problem the small error due to the first order approximation to ΔT probably did not warrant the refinements obtained by the calculation of the true solution. It should also be pointed out that this example illustrates a procedure which may be used to determine the true solution by using the first order approximation as a starting point.

SUMMARY

A control system sensitivity index has been introduced which utilizes either the mean square value or the integral square value of the error in the system response due to plant parameter variations. This measure of sensitivity is simplified by expressing the change in the system transfer function by a first order approximation to the actual change. This simplification results in an expression for the system sensitivity index which is a function of the numerator coefficients of the Bode sensitivity function and the plant parameter deviations from their nominal values. The analysis of control system sensitivity according to this criterion is then easily carried out by applying Parseval's Theorem and evaluating the integral of a frequency domain function from published tables [9].

A synthesis technique is developed for minimizing the system sensitivity index under worst case conditions with respect to numerator coefficients of the Bode sensitivity function. A design example is used to illustrate the synthesis of a control system for minimum sensitivity and to compare the sensitivity of this design with a number of other configurations. This comparison shows that the synthesis technique does significantly decrease the system sensitivity. The errors incurred due to the first order approximation are investigated for this design example and it is shown the errors are not significant. Finally a method is developed which may be used to determine the true solution by using the first order appoximation as a starting point for an iterative procedure.

ACKNOWLEDGEMENT

The author wishes to express his gratitude to Professors J. B. Crux, Jr., W. R. Perkins, and M.E. VanValkenburg for their invaluable guidance and encouragement throughout the period of research for this paper.

LIST OF REFERENCES

1. Blostein, M. L., "Some Bounds of Sensitivity in RLC Networks" Proc. First Allerton Conf. on Circuit and System Theory, pp. 488-501; November, 1963.
2. Bode, H. W., Network Analysis and Feedback Amplifier Design. Van Nostrand, New York, N. Y., 1945.
3. Goldstein, A. J. and Kuo, F. F., "Multiparameter Sensitivity" IRE Trans. on Circuit Theory, Vol. CT-8, No. 2, pp 177-178; June, 1961.
4. Gonzales, R. L., "Control System Synthesis to Minimize Multiparameter Sensitivity," Coordinated Science Laboratory Report R-312, Contract No. DA 28 043 AMC 00073(E) University of Illinois, August, 1966.
5. Hakimi, S. L. and Cruz, J. B., Jr., "Measures of Sensitivity for Linear Systems with Large Multiparameter Variation," IRE Wescon Convention Record, Part 2, pp. 109-115; August, 1960.
6. Hooke, R. and Jeeves, T. A., "Direct Search Solution of Numerical and Statistical Problems," J. Assoc. Comp. Mach., pp. 212-229, April, 1961.
7. Mazer, W. M., "Specifications of the Linear Feedback System Sensitivity Function," IRE Trans. on Automatic Control, Vol. AC-5, pp. 85-93, June, 1960.

8. Mikulski, J. H., "A Correlation Between Classical and Pole-Zero Sensitivity," Technical Note No. 55, Contract No. AF 49(638)-63, Electrical Engineering Research Labs., University of Illinois, 1959.
9. Newton, G. C., Gould, L. A. and Kaiser, J. F., Analytical Design of Linear Feedback Controls, John Wiley and Sons, Inc., New York, N.Y., 1957.
10. Remec, M. J., "Application of Multiparameter Sensitivity in Linear Systems," Doctoral Dissertation, University of Illinois, 1963.
11. Rutman, R. S., "Statistical Synthesis of Control Systems which are Optimal with Respect to Insensitivity to Variation of Parameters," Izv. Akad. Nauk SSSR, Otd. Teklin. Nauk: Kibernetika (Moscow), No. 2, pp. 188-199, 1965.

SENSITIVITY INTEGRALS

SYDNEY R. PARKER
U. S. Naval Postgraduate School
Monterey, California

ABSTRACT: In this paper the effects of incremental parameter variations are treated in a broad sense as equivalent to perturbation signals. First, logarithmic sensitivity is expressed directly in terms of the polynomials of a transfer function and the equivalent input perturbation signal defined. As a corollary a general analog computer simulation for sensitivity studies is presented. In order to compare the relative sensitivities of different parameters it is convenient to have an overall measure of sensitivity which is independent of either time or frequency. To this end the area of the equivalent perturbation signal, and the areas of the incremental response characteristics in the time and sinusoidal frequency domains (called sensitivity integrals) are investigated. The sensitivity squared integral, given by the area of the equivalent input perturbation signal squared, is also presented. Various interrelationships of these integrals and generalized results are presented. Application is also made to several specific examples.

INTRODUCTION

The Equivalent Input Perturbation Signal

It is known[1] that for any linear system the transfer function, T(s,x), relating a response to an input function can be expressed in the following bilinear form in terms of a real parameter, x, of the system and the polynomials A(s), B(s), C(s), D(s).

$$T(s,x) = \frac{P(s,x)}{Q(s,x)} = \frac{A(s) + xB(s)}{C(s) + xD(s)} \qquad (1)$$

where

$$P(s,x) = A(s) + xB(s) \qquad (2)$$

$$Q(s,x) = C(s) + xD(s) \qquad (3)$$

From (2) and (3) it follows that

$$A(s) = P(s,0) \qquad (4)$$

$$B(s) = \frac{1}{x}\left[P(s,x) - P(s,0)\right] \qquad (5)$$

$$C(s) = Q(s,0) \qquad (6)$$

$$D(s) = \frac{1}{x}\left[Q(s,x) - Q(s,0)\right] \qquad (7)$$

With the aid of the foregoing equations the logarithmic sensitivity function may be expressed directly in terms of the polynomials of the original transfer function.

$$S(s,x) = \frac{\partial \ell n T(s,x)}{\partial \ell n x} = \frac{x}{T(s,x)} \frac{\partial T(s,x)}{\partial x}$$

$$= \left[\frac{Q(s,0)}{Q(s,x)} - \frac{P(s,0)}{P(s,x)}\right] \tag{8}$$

If the variable x changes by an incremental amount, Δx, then as an approximation given by the first terms of a Taylor series, T(s,x) changes by an incremental amount given by

$$\begin{aligned} \Delta T(s,x) &= T(s,x + \Delta x) - T(s,x) \\ &= \frac{\partial T(s,x)}{\partial x} \Delta x = \left(\frac{\Delta x}{x}\right) T(s,x)S(s,x) \end{aligned} \tag{9}$$

For a given input function, R(s), the systems normal response is given by, $Y(s,x) = R(s)T(s,x)$. The change in the response, $\Delta Y(s,x)$, caused by the increment Δx is given by, $\Delta Y(s,x) = R(s) \Delta T(s,x)$. Substituting (9) for $\Delta T(s,x)$ leads to the result

$$\Delta Y(s,x) = \left(\frac{\Delta x}{x}\right) R(s)T(s,x)S(s,x) \tag{10}$$

Eq. (10) indicates that, $\Delta Y(s,x)$, the increment in the response, may be obtained by calculating the response due to an equivalent input perturbation signal, $\Delta R(s)$, applied to the original transfer function, T(s,x), where

$$\Delta R(s) = \left(\frac{\Delta x}{x}\right) R(s)S(s,x) \tag{11}$$

Simulation

Eq. (10) may be represented in block diagram form as drawn in Fig. 1, where S(s,x) has been represented according to (8). From (3), polynomials Q(s,x) and Q(s,0) can be written in expanded form as follows

$$Q(s,x) = (c_o + xd_o) + (c_1 + xd_1)s + \ldots + (c_{m-1} + xd_{m-1})s^{m-1} + s^m \tag{12}$$

$$Q(s,0) = c_o + c_1 s + \ldots + c_{m-1} s^{m-1} + s^m \tag{13}$$

Simulation of the transfer function Q(s,0)/Q(s,x) may be accomplished on an analog computer as illustrated in Fig. 2. A similar simulation is possible for the transfer function P(s,0)/p(s,x), so that interconnection according to Fig. 1 yields an analog computer method for obtaining the change in the response, $\Delta y(t,x)$, directly as a time function trace for any input signal. Obviously the simulation can be extended to yield the output variation function for increments in several independent variables at the same time. For the special case where T(s,x) has poles only and x is taken as a coefficient of the denominator polynomial, the simulation of Fig. 1 can be shown to reduce to one suggested by R. Tomovic[2] for obtaining sensitivity coefficients.

Sinusoidal Steady-State

In the sinusoidal steady-state, $s = j\omega$

$$T(j\omega,x) = \left| T(\omega,x) \right| e^{j\theta(\omega,x)} \tag{14}$$

and

$$\ell n T(j\omega,x) = \ell n \left| T(\omega,x) \right| + j\theta(\omega,x) \tag{15}$$

From (15) and the definition of $S(j\omega,x)$ as given by (8), it follows that

$$\mathrm{Re}\left[S(j\omega,x)) \right] = \frac{\partial \ell n \left| T(\omega,x) \right|}{\partial \ell n x} \tag{16}$$

$$\mathrm{Im}\left[S(j\omega,x) \right] = \frac{\partial \theta(\omega,x)}{\partial \ell n x} \tag{17}$$

Using (16) and (17) it can be shown that if x changes by an increment Δx, the corresponding incremental changes, $\Delta \ell n \left| T(\omega,x) \right|$, and, $\Delta \theta(\omega,x)$, are given by

$$\frac{\Delta \ell n \left| T(\omega,x) \right|}{\left(\frac{\Delta x}{x} \right)} = \mathrm{Re}\left[S(j\omega,\ x) \right] \tag{18}$$

$$\frac{\Delta \theta\ (\omega,x)}{\left(\frac{\Delta x}{x} \right)} = \mathrm{Im}\left[S(j\omega,x) \right] \tag{19}$$

Zero-pole Sensitivities

Finally, the polynomial ratios $Q(s,0)/Q(s,x)$ and $P(s,0)/P(s,x)$ which compose $S(s,x)$ may be related to the zero-pole sensitivities of the transfer function as follows. Consider

$$P(s,x) = K(x) \prod_{i=1}^{m} (s + z_i(x)) \tag{20}$$

$$Q(s,x) = \prod_{i=1}^{n} (s + p_i(x)) \tag{21}$$

where the zero-pole locations, $-z_i(x)$ and $-p_i(x)$, as well as the constant $K(x)$ are assumed to be functions of the variable x. From (3) and (7) it can be shown that

$$\frac{\partial \ell n Q(s,x)}{\partial x} = \frac{1}{x} \left[1 - \frac{Q(s,0)}{Q(s,x)} \right] \tag{22}$$

Taking the logarithm of (21), differentiating with respect to x, and equating the result to (22) yields

$$\frac{Q(s,0)}{Q(s,x)} = 1 - \sum_{i=1}^{n} \frac{x}{s + p_i(x)} \frac{dp_i(x)}{dx} \tag{23}$$

In a similar way it can be demonstrated that

$$\frac{P(s,0)}{P(s,x)} = 1 - \frac{x}{K(x)} \frac{dK(x)}{dx} - \sum_{i=1}^{m} \frac{x}{s + z_i(x)} \frac{dz_i(x)}{dx} \tag{24}$$

SENSITIVITY INTEGRALS

The foregoing expressions for sensitivity and incremental variations lead to functions of the parameter x and either s, t, or ω. From an engineering point of view it would be convenient to have an overall measure of sensitivity which is independent of either time or frequency. Such a figure would be useful for comparing the relative sensitivities of different parameters in a system as well as presenting the possibility of including a sensitivity criteria for purposes of design or optimization.

The Phase Integral

Consider first the area of the equivalent input perturbation signal, $\Delta r(t,x)$, relative to the area of the input signal, r(t).

The total area of the input signal is given by the integral

$$I_i = \lim_{t\to\infty} \int_0^t r(t)dt = \lim_{t\to\infty} \mathcal{L}^{-1}\left[\frac{R(s)}{s}\right] = \lim_{s\to 0} \frac{s}{s} R(s) = R(0) \tag{25}$$

where use is made of the fact that $\int_0^t f(t)dt = \mathcal{L}^{-1}\left[\frac{F(s)}{s}\right]$ and the Final Value theorem applied. With the aid of (11) the total area of the equivalent input perturbation signal is given by

$$\Delta I_i = \lim_{t\to\infty} \int_0^t \Delta r(t)dt = \left(\frac{\Delta x}{x}\right) \lim_{t\to\infty} \mathcal{L}^{-1}\left[\frac{R(s)S(s,x)}{s}\right]$$

$$= \left(\frac{\Delta x}{x}\right) \lim_{s\to 0} \frac{s}{s} R(s)S(s,x) = \left(\frac{\Delta x}{x}\right) R(0)S(0,x) \tag{26}$$

If $I_i \neq 0$ the ratio, I(x), of (26) to (25) exists[3] and is given by

$$I(x) = \frac{\Delta I_i}{I_i} = \left(\frac{\Delta x}{x}\right) S(0,x) \tag{27}$$

When $I_i = 0$, it follows from (26) that $\Delta I_i = 0$ provided $S(0,x) \neq \infty$. When $R(0) = 0$ and $S(0,x) = \infty$, ΔI_i may be evaluated by calculating the limit of the product R(s)S(s,x) as indicated by (26).

Alternately I(x) may be obtained by considering the area of the change in the response due to the increment Δx. Prior to the parameter change the area of the response is given by

$$I_o = \lim_{t\to\infty} \int_0^t y(t,x)dt = \lim_{t\to\infty} \mathcal{L}^{-1}\left[\frac{R(s)T(s,x)}{s}\right]$$

$$= \lim_{s\to 0} \frac{s}{s} R(s)T(s,x) = R(0)T(0,x) \tag{28}$$

After the increment Δx, the area of the change in the response is, with the aid of (10), given by

$$\Delta I_o = \lim_{t\to\infty} \int_o^t \Delta y(t,x)dt = (\frac{\Delta x}{x}) \lim_{t\to\infty} \mathcal{L}^{-1}\left[\frac{R(s)T(s,x)S(s,x)}{s}\right]$$

$$= (\frac{\Delta x}{x}) \lim_{s\to 0} \frac{s}{s}\left[R(s)T(s,x)S(s,x)\right] = (\frac{\Delta x}{x})R(0)T(0,x)S(0,x) \qquad (29)$$

If $I_o \neq 0$, the ratio of (29) to (28) exists[3] and is given by

$$\frac{\Delta I_o}{I_o} = (\frac{\Delta x}{x})\ S(0,x) = I(x) \qquad (30)$$

When $I_o = 0$ it follows from (29) that $\Delta I_o = 0$ provided $S(0,x) \neq \infty$. When $I_o = 0$ and $S(0,\infty) = \infty$, ΔI_o may be evaluated by calculating the limit of the product $R(s)T(s,x)S(s,x)$ as indicated by (29).

Now consider sinusoidal steady-state and the area of the incremental phase characteristic, $I_\theta(x,\omega)$ plotted on a logarithmic frequency scale, defined by the integral

$$I_\theta(\omega,x) = \int_{-\infty}^{\infty} \Delta\theta(\omega,x)d(\ell n\omega) = \int_{-\infty}^{\infty}\Delta\theta(\omega,x)\frac{d\omega}{\omega} = (\frac{\Delta x}{\Delta})\int_{-\infty}^{\infty} \mathrm{Im}\left[S(j\omega,x)\right]\frac{d\omega}{\omega} \qquad (31)$$

where use has been made of (19). From the Reactance Integral Theorem[1] it is known that if $H(j\omega) = R(\omega) + jX(\omega)$ and $\lim_{\omega\to\infty} H(j\omega) = 0$, then

$$\int_{-\infty}^{\infty} X(\omega)\frac{d\omega}{\omega} = -\frac{\pi}{2}\lim_{s\to 0} H(s) \qquad (32)$$

Applying (32) to (31) yields the result

$$I_\theta(x,\omega) = -\frac{\pi}{2}(\frac{\Delta x}{x})\lim_{s\to 0} S(s,x) = -\frac{\pi}{2}(\frac{\Delta x}{x})S(0,x) = -\frac{\pi}{2} I(x) \qquad (33)$$

Eqs. (27),(30), and (33) enable the following theorem to be stated.

Theorem. For an incremental parameter change, Δx, the product of $(\frac{\Delta x}{x})$ times the logarithmic sensitivity function evaluated at $s = 0$ is equal to (a) the ratio of the area of the equivalent input perturbation signal to the area of the input signal, provided the latter area is non-zero. (b) The fractional change in the system's response to any input signal, provided the area of the response is non-zero. (c) $(-\frac{2}{\pi})$ times the change in area of the sinusoidal steady-state phase characteristic plotted on a logarithmic frequency scale.

When $S(0,x) = 0$, (33) indicates that the increment Δx causes zero net change in the area of the sinusoidal steady-state phase characteristic plotted versus $\ell n\omega$. Similarly, when $S(0,x) = 0$, (29) indicates that Δx produces zero net change in the area of any time response characteristic which initially has finite non-zero area. If the area of the response is zero and $S(0,x)$ is finite there will also be zero net change in area produced by Δx. $S(0,x)$ may be calculated directly from (8)

$$S(0,x) = \frac{Q(0,0)}{Q(0,x)} - \frac{P(0,0)}{P(0,x)} \tag{33a}$$

$$= \frac{c_o}{c_o + xd_o} - \frac{a_o}{a_o + xb_o} \tag{33b}$$

where $P(0,x) = a_o + xb_o$ and $Q(0,x) = c_o + xd_o$. From (33b) it is easy to show that $S(0,x)$ will be identically zero when $a_o d_o = c_o b_o$.

Example. Type N Unity Feedback Systems

It shall now be demonstrated that $S(0,x) = 0$ for the closed loop transfer function of any type $N > 0$ unity feedback control system which has an open loop transfer function given by

$$G(s) = \frac{K \prod_{i=1}^{m} (1 + s\tau_{zi})}{s^N \prod_{i=1}^{n} (1 + s\tau_{qi})} \tag{34}$$

The closed loop transfer function is given by $T(s,x) = P(s,x)/Q(s,x)$ where

$$P(s,x) = K \prod_{i=1}^{m} (1 + s\tau_{zi}) \tag{35}$$

$$Q(s,x) = s^N \prod_{i=1}^{n} (1 + s\tau_{qi}) + K \prod_{i=1}^{m} (1 + s\tau_{zi}) \tag{36}$$

The parameter x need not be specified at first but will be either K, τ_{qi} or τ_{zi}. From (35) and (36) it follows that for $N > 0$

$$P(0,x) = Q(0,x) = K \tag{37}$$

$$\begin{aligned} P(0,0) = Q(0,0) &= K \text{ when } x \neq K \\ &= 0 \text{ when } x = K \end{aligned} \tag{38}$$

When $x \neq K$, it follows that $P(0,0)/P(0,x) = Q(0,0)/Q(0,x) = 1$, so that from (33a), $S(0,x) = 0$. It is interesting to note from (23) and (24) with $s = 0$, that the foregoing leads to the following relationship for the zero-pole sensitivities of the closed loop system.

$$\sum_{i=1}^{n} \frac{x}{p_i(x)} \frac{dp_i(x)}{dx} = \sum_{i=1}^{m} \frac{x}{z_i(x)} \frac{dz_i(x)}{dx} = 0 \tag{39}$$

When $x = k$, $P(0,0)/P(0,x) = Q(0,0)/Q(0,x) = 0$ and again $S(0,x) = 0$. For this case from (23) with $s = 0$, it follows that

$$\sum_{i=1}^{n} \frac{K}{p_i(K)} \frac{dp_i(K)}{dK} = 1 \tag{40}$$

Finally, when $N = 0$ it can be demonstrated that $S(0,x) = 0$ when $x \neq K$, and $S(0,x) = (1 + K)^{-1}$ when $x = K$.

The Magnitude Integral

It is also of interest to investigate the area of the change in the sinusoidal steady-state magnitude characteristic as given by (18). Consider the integral

$$\Delta I_M(\omega, x) = \int_0^\infty \Delta \ln \left| T(\omega,x) \right| d\omega = \frac{\Delta x}{x} \int_0^\infty \Re\left[S(j\omega,x)\right] d\omega \qquad (41)$$

From the Resistance Integral Theorem[1] it can be shown that if $H(j\omega) = R(\omega) + jX(\omega)$ and $\lim_{\omega \to \infty} H(j\omega) = 0$ then

$$\int_0^\infty R(\omega) d\omega = \frac{\pi}{2} \lim_{s \to \infty} sH(s) \qquad (42)$$

Applying (42) to (41) yields the result

$$\Delta I_M(\omega,x) = \left(\frac{\Delta x}{x}\right) \frac{\pi}{2} \lim_{s \to \infty} sS(s,x) \qquad (45)$$

Eq. (43) is also easily evaluated from (8). The use of this integral is illustrated in the next section.

Sensitivity Squared Integral

Since $S(0,x)$ is often zero it is not always a convenient criterion for comparing the sensitivities of different parameters in a system. This suggests the possibility of investigating the areas of the incremental error signals squared.

Using an ingenious numerical integration technique, W. M. Mazer[5] has applied the criterion of minimum mean square error to the incremental response of a linear feedback system for one cycle of a fixed frequency sinusoidal driving function. A simpler and less restrictive approach is to investigate the area of the equivalent input perturbation signal squared. Thus, the following integral is defined

$$I^2(x) = \int_0^\infty \left[\Delta r(t,x)\right]^2 dt = \int_{-\infty}^\infty \left[\Delta r(t,x)\right]^2 dt \qquad (44)$$

where it is assumed that $\Delta r(t,x)$ is zero for $t < 0$. Applying Parceval's theorem[4] to (44), with the aid of (11), yields

$$I^2(x) = \left(\frac{\Delta x}{x}\right)^2 \int_{-j\infty}^{+j\infty} \left[R(s)S(s,x)\right]\left[R(-s)S(-s,x)\right] ds \qquad (45)$$

When $R(s) = 1$, $r(t) = \delta(t)$, and (45) becomes

$$I_\delta^2(x) = \left(\frac{\Delta x}{x}\right)^2 \int_{-j\infty}^{+j\infty} S(s,x)S(-s,x) ds \qquad (46)$$

Integrals of the type of (45) and (46) have been tabulated in the literature[4] and form the basis of analytical design proceedures. A similar approach is suggested for system sensitivity studies since (46) represents the integral-square of the equivalent input perturbation signal which produces the same variation in the impulsive response as the parameter change Δx. In particular (46) can form a basis of comparison for the affects of different parameter variations upon system performance. The approach may be illustrated by a simple example. Consider a second order positional servomechanism subject to a load disturbance torque $T_L(t)$ applied to the output shaft. The transfer function relating the output angle to the applied torque is given by

$$\frac{\theta_o(s)}{T_L(s)} = \frac{1}{Js^2 + Fs + K} \tag{47}$$

where J, F, and K are moment of inertia, viscous damping, and gain respectively. It is desired to compare the sensitivity of the impulsive response, when $T_L(t) = \delta(t)$, for incremental variations of F and K. From the polynomials of (47) it follows that when $x = F$, $P(s,F) = 1$; $P(s,0) = 1$; $Q(s,F) = Js^2 + Fs + K$; $Q(s,0) = Js^2 + K$. Substituting into (8) yields

$$S(s,F) = \frac{-Fs}{Js^2 + Fs + K} \tag{48}$$

Similarly it can be shown that

$$S(s,K) = \frac{-K}{Js^2 + Fs + K} \tag{49}$$

Substituting (48) and (49) into (46) and performing the required integrations using the tables in reference (3), yields

$$I_\delta^2(F) = \frac{F}{2J}\left(\frac{\Delta F}{F}\right)^2 = \zeta\omega_n\left(\frac{\Delta F}{F}\right)^2 \tag{50}$$

$$I_\delta^2(K) = \frac{K}{2F}\left(\frac{\Delta K}{K}\right)^2 = \frac{\omega_n}{4\zeta}\left(\frac{\Delta K}{K}\right)^2 \tag{51}$$

where $\omega_n = \sqrt{K/J}$ and $\zeta = F/2\sqrt{KJ}$, the undamped resonant frequency and damping ratio respectively. Eqs. (50) and (51) indicate that both sensitivity squared integrals are proportional to ω_n. For $\zeta > \frac{1}{2}$ a fractional variation of F is seen to have a greater effect upon its sensitivity squared integral than the identical fractional variation of K has upon its integral. The implied conclusion is that for $\zeta > \frac{1}{2}$, fractional variations of F have a greater overall affect upon the impulsive response than equal fractional variations of K. The reverse of this situation holds true for $\zeta < \frac{1}{2}$. It is also interesting to evaluate the change in magnitude integral of (43) for the present example. From (43), (48), and (49) it follows that

$$\Delta I_M(\omega,F) = \frac{-\pi}{2}\frac{F}{J}\left(\frac{\Delta F}{F}\right) = -\pi\zeta\omega_n\left(\frac{\Delta F}{F}\right) \qquad (52)$$

$$\Delta I_M(\omega,K) = 0 \qquad (53)$$

Eqs. (52) and (53) indicate that variation of K does not affect the net area of the log magnitude versus ω characteristic, whereas variation of F affects the net area of this characteristic in a manner similar to its affect upon its sensitivity squared integral, as can be seen by comparing (52) with(50). It remains to apply the foregoing sensitivity integrals to more sophisticated examples and to ascertain whether the technique can be incorporated meaningfully into design specifications and proceedures.

REFERENCES

1. Bode, H. W., "Network Analysis and Feedback Amplifier Design", D. Van Nostrand Co., New York, 1945. Sects. 1.7, 13.4 and 13.7.

2. Tomovic, R., "Sensitivity Analysis of Dynamic Systems", McGraw-Hill Book Col, New York, 1963. Sect. 3.5.4.

3. Kaplan, W, "Advanced Calculus", Addison-Wesley Publishing Co., Reading, Mass, 1957, Section 9-7, Theorem 12.

4. Newton, G. C., Gould, L. A., Kaiser, J. F., "Analytic Design of Linear Feedback Controls", John Wiley & Sons, Inc., New York, 1957. Sect. 2.2 and Appendix E.

5. Mazer, W. M., "Specification of the Linear Feedback System Sensitivity Function", IRE Transactions on Automatic Control, June 1960, pp 85-93.

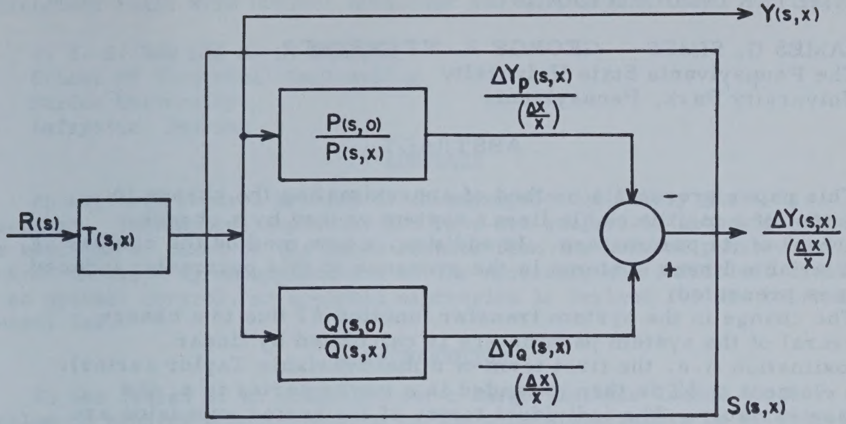

Y(s,x)
$\frac{\Delta Y_P(s,x)}{(\frac{\Delta x}{x})}$
$\frac{P(s,0)}{P(s,x)}$
R(s)
T(s,x)
-
+
$\frac{\Delta Y(s,x)}{(\frac{\Delta x}{x})}$
$\frac{Q(s,0)}{Q(s,x)}$
$\frac{\Delta Y_Q(s,x)}{(\frac{\Delta x}{x})}$
S(s,x)

FIG. I BLOCK DIAGRAM FOR GENERATING THE INCREMENTAL RESPONSE CAUSED BY THE PARAMETER CHANGE, ΔX.

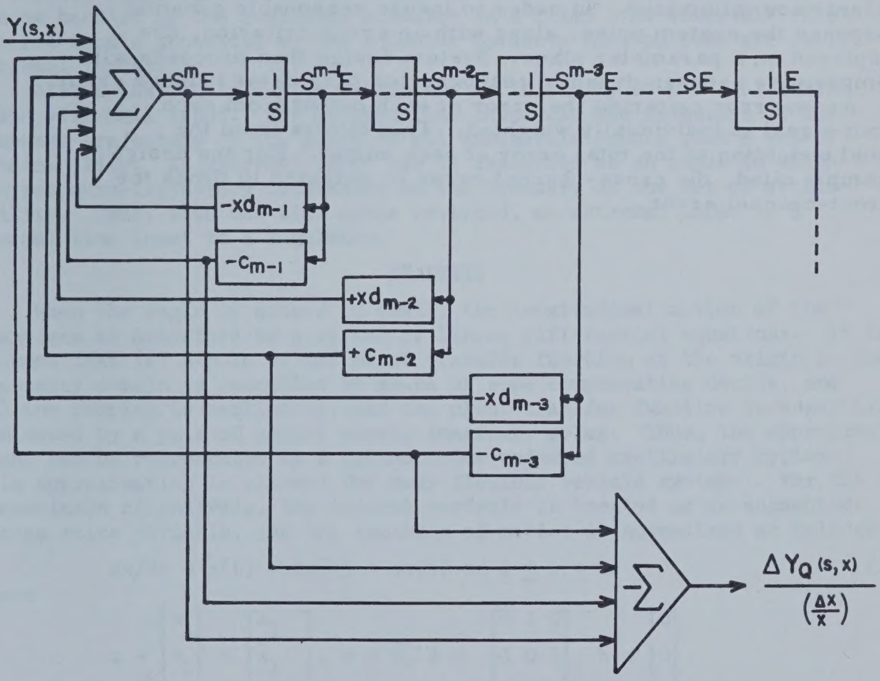

Y(s,x)
$-\sum$
$+S^m E$
$-\frac{1}{S}$
$-S^{m-1}E$
$-\frac{1}{S}$
$+S^{m-2}E$
$-\frac{1}{S}$
$-S^{m-3}E$
$-SE$
$-\frac{1}{S}$
E
$-xd_{m-1}$
$-c_{m-1}$
$+xd_{m-2}$
$+c_{m-2}$
$-xd_{m-3}$
$-c_{m-3}$
$-\sum$
$\frac{\Delta Y_Q(s,x)}{(\frac{\Delta x}{x})}$

FIG. 2 SIMULATION OF THE TRANSFER FUNCTION,

$$\frac{Q(s,0)}{Q(s,x)} = \left(\frac{E}{Q(s,x)}\right)\left(\frac{Q(s,0)}{E}\right)$$

DESIGN OF MULTIVARIABLE LINEAR SYSTEMS BY MINIMIZATION OF PARAMETER INDUCED ERROR

JAMES G. SIMES GEORGE A. ETZWEILER
The Pennsylvania State University
University Park, Pennsylvania

ABSTRACT

This paper presents a method of approximating the change in the output of a multivariable linear system caused by a change in several of its parameters. In addition, a new method for comparing multivariable linear systems in the presence of this parameter induced error is presented.

The change in the system transfer function δT due to a change in several of the system parameters is calculated by linear approximation (i.e. the first term of a multivariable Taylor series). Each element of δT is then expanded in a power series in s, the Laplace variable. The individual terms of the series expansion are referred to as "Parameter Error Coefficients". If the input is in the form of a polynomial in t, then this series expansion of δT times the input vector M gives an exact expression for the steady state parameter induced error, at least within the linear approximation. As a design procedure the first few parameter error coefficients are minimized. In order to insure reasonable dynamic response the system poles, along with an error criterion, are displayed on a parameter plane. System design then proceeds with compromise between dynamic response and parameter induced error.

As an error criterion the error at each output from each input signal is individually weighted. This differs from the usual weighting of the total error at each output. For the design example cited, the cross-channel error is weighted 10 times the direct-channel error.

AN APPLICATION OF BOUNDED PHASE-COORDINATE CONTROL THEORY TO AN UNDAMPED OSCILLATORY PLANT WITH CONTROL AMPLITUDE AND RATE SATURATIONS

J. Y. S. LUH AND J. S. SHAFRAN
School of Electrical Engineering
Purdue University
Lafayette, Indiana

ABSTRACT

An extremal control problem of an undamped oscillatory plant is discussed. The control amplitude and rate are subject to saturation limits. The analysis is based on the known results from the bounded phase-coordinate control theory. By an application of the necessary and sufficient conditions of an optimal control, an analytic expression is derived for the extremal control law.

INTRODUCTION

In the design of an autopilot for a large flexible launch booster, the problem of bending moments related to the wind disturbances is of relative importance. When the motion-controlling actuator has saturation limits on both position and rate, the design problem is quite involved. As a rule, the autopilot should be capable of maneuvering the actuator in a most efficient manner while encountering the worst wind disturbance. Therefore, an initial investigation is the determination of the worst disturbance that can be handled by the available actuator in a fixed time interval. The control inputs generated by the actuator to serve this purpose are called extremal inputs.

Assume that the plant (the booster) is initially at its equilibrium state with null input. For a fixed time interval, the endpoints of the response-trajectory of the plant to all admissible inputs form a set of attainability. An extremal input can be described as an input to which the response-trajectory terminates at the boundary of the set of attainability. Thus, with the time sense reversed, an extremal input is a minimal time input to a regulator.

ANALYSIS

When the angle of attack is small, the longitudinal motion of the plant can be described by a system of linear differential equations. It is assumed that (a) a pole of the plant transfer function at the origin in the frequency domain is cancelled by means of some compensating device, and (b) the damping is negligible, and the plant transfer function is essentially dominated by a pair of almost purely imaginary poles. Thus, the approximate plant can be represented by a second order undamped oscillatory system. This approximation is allowed for many flexible vehicle systems. For the convenience of analysis, the control variable is treated as an augmented-system state variable, and the equation of motion is normalized as follows:

$$dx/dt = \mathring{x}(t) = Ax(t) + bv(t) \text{ on } 0 \leq t \leq t_1 \qquad (1)$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ \mathring{x}_1 \\ u(t) \end{bmatrix}, \; v = \mathring{u}, \; A = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

x_1 = normalized plant position, $u(t)$ = normalized thrust deflection with $|u| \leq 1$, $v(t)$ = normalized thrust deflection rate with $|v| \leq D$. A typical value of D is one-fifth of the period of the oscillation $(= 2.5/\pi)$ which allows the extremal control to enter upon and exit from its bound once every half cycle of the oscillation. The three-dimensional coordinate system is so chosen that the origin is an equilibrium state. It is required to determine an extremal v(t) which steers the system (1) from the origin to a furthest point x in a given direction in a fixed time t. This is a bounded phase-coordinate control problem (in the sense $|x_3| = |u| \le 1$), and all known results on this type of problem apply.

Gamkrelidge [1] showed that, in order to represent the extremal v as a multiple of the signum of an adjoint solution for the bounded phase-coordinate control problem, the modified adjoint vector p(t) must satisfy the relation

$$\mathring{p}(t) = \begin{cases} -A'p(t), & \text{if } |x_3| < 1 \\ -\tilde{A}'p(t), & \text{if } |x_3| = 1 \end{cases} \tag{2}$$

where ( )' = transpose of ( ) and

$$\tilde{A} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} .$$

Thus, $v(t) = 2.5 \operatorname{sgn} [p(t)'b]/\pi = 2.5 \operatorname{sgn} p_3(t)/\pi$ in which

$$\operatorname{sgn} p_3 = \begin{cases} +1, & \text{if } p_3 > 0 \\ 0, & \text{if } p_3 = 0 \\ -1, & \text{if } p_3 < 0 \end{cases}$$

Furthermore, p_3 is allowed certain jump discontinuities at endpoints of intervals where $|x_3| = 1$; and $p_3 = 0$ whenever $|x_3| = 1$. As a result, the solution of the system (2) is

$$p_1(t) = p_1(o) \cos t + p_2(o) \sin t,$$
$$p_2(t) = - p_1(o) \sin t + p_2(o) \cos t,$$
$$p_3(t) = \begin{cases} -p_1(o) \cos t - p_2(o) \sin t + k, & \text{if } |x_3| < 1, \\ 0 & , \text{if } |x_3| = 1, \end{cases}$$

where the value of constant k in p_3 depends on the initial conditions of p_1 and p_2, and on the interval in which $|x_3| < 1$. Let

$$K(t) = \{x(t) | x(t) = \int_0^t e^{A(t-s)} bv(s)\, ds,\ |x_3| \le 1,\ |v| \le \pi/2.5\}$$

be a set of attainability at time t. The transversality condition implies that for every unit vector η in R^3 there is a state vector x corresponding to a point on the boundary of K(t) for a fixed t such that the projection, P, of x onto η is a maximum. Moreover, the corresponding unique admissible extremal v(s), which maximizes P, steers the system (1) from the origin to the furthest point x in the direction η in a fixed time t. Russell's sufficiency condition [2] shows that η is the adjoint vector at time t, and v is the minimal-time control. Let

$$\eta = \begin{bmatrix} \cos \theta \ \cos \phi \\ \sin \theta \ \cos \phi \\ \sin \phi \end{bmatrix}, \ |\theta| \le \pi,\ |\phi| < \pi/2 ,$$

then

$$P = \cos \phi \int_0^t [\cos \theta - \cos(t + \theta - s) + \tan \phi]\, v(s)\, ds .$$

The extremal v which maximizes P can be determined by inspection for every fixed θ, $\emptyset$ and t. Since $P(s;\ t,\ \theta,\ \emptyset) = -\ P(s;\ t,\ \pi + \theta,\ -\ \emptyset)$, it suffices to consider only half of the range of θ. Figure 1 shows a typical case of $3\pi/5 \leq \theta \leq 4\pi/5$, with $11\pi/5 < t + \theta < 14\pi/5$. It was found that, in general, the extremal v reaches zero and leaves zero as often as the length of t permits. Denote the time at which such events occur by τ_i, $i = 1,2,..,2N$, and let $\tau_o = 0$, $\tau_{2N+1} = t$. Furtheremore, let

$$|x_3(s)| < 1 \Rightarrow dp_3/ds = d\psi_3/ds, \text{ if } \tau_{2i} \leq s < \tau_{2i+1},\ i = 0,\ldots,\ N;$$

$$|x_3(s)| = 1 \Rightarrow p_3(s) = 0, \text{ if } \tau_{2j+1} \leq s < \tau_{2j+2},\ j = 0,\ldots,\ N-1;$$

where ψ_3 is a component of ψ satisfying $d\psi/ds = -\ A'\psi$. It follows that choosing $p_3(s)$ to be continuous at τ_{2i+1} requires $p_3(\tau_{2i+1}) = 0$, and hence

$$p_3(s) = \psi_3(s) - \psi_3(\tau_{2i+1}) \text{ for } \tau_{2i} \leq s < \tau_{2i+1},\ i = 0,\ldots,\ N\ .$$

Thus,

$$\frac{p_3(s;\ t,\ \theta,\ \emptyset)}{\cos\ \emptyset} = \begin{cases} \cos(t - \tau_{2i+1} + \theta) - \cos(t - s + \theta), & \text{if } \tau_{2i} \leq s < \tau_{2i+1}, \\ 0 & \text{, if } \tau_{2i+1} \leq s < \tau_{2i+2}, \\ \cos\ \theta - \cos(t - s + \theta) + \tan\ \emptyset + \delta(\tan\ \emptyset_o - \tan\ \emptyset), & \\ & \text{if } \tau_{2N} \leq s < t, \end{cases}$$

where $i = 0,\ 1,\ldots,\ N - 1$,

$$\delta = \begin{cases} 0, & \text{if } |x_3(t)| = |u(t)| < 1, \\ 1, & \text{if } |x_3(t)| = |u(t)| = 1, \end{cases}$$

and

$\emptyset_o$ = direction limit for η at which p_3 has a jump discontinuity.

Since $\cos\ \emptyset > 0$ for $|\emptyset| < \pi/2$, the extremal v(s) than can be expressed as

$$v(s) = \frac{2.5}{\pi}\ \text{sgn}[p_3(s;\ t,\ \theta,\ \emptyset)/\cos\ \emptyset]$$

$$= \frac{2.5}{\pi}\ \text{sgn}[p_3(s;\ t,\ \theta,\ \emptyset)]\ .$$

The formulas for parameters τ_i, $i = 1,\ldots,\ 2N$, δ and $\emptyset_o$ are determined for all possible choices of θ, $\emptyset$, and t, and are presented in two charts. To use these charts, first locate the Case Number from Chart I for the appropriate ranges of t and θ. Then on Charts II, for every Case Number and every range of $\emptyset$, a set of parameters are given.

CONCLUSION

The analysis presented in this paper is a direct application of the optimal control theory. The method is derived from the necessary and sufficient conditions of an optimal control. This investigation reveals the structure of the extremal control input, which oscillates in accordance with the oscillation of the controlled plant (the booster). Furthermore, the analysis yields an analytic solution for the extremal control law which is an explicit time function.

REFERENCES

1. Pontryagin, L. S. et al., The Mathematical Theory of Optimal Processes, Interscience, 1962.
2. Luh, J. Y. S., D. L. Russell, E. B. Lee, Bounded Phase-Coordinate Control, Final Technical Report to NASA Contract NASw986, Honeywell Document No. 12006-FRI, July 1965. Also NASA Contractor Report NASA CR-470, May 1966.

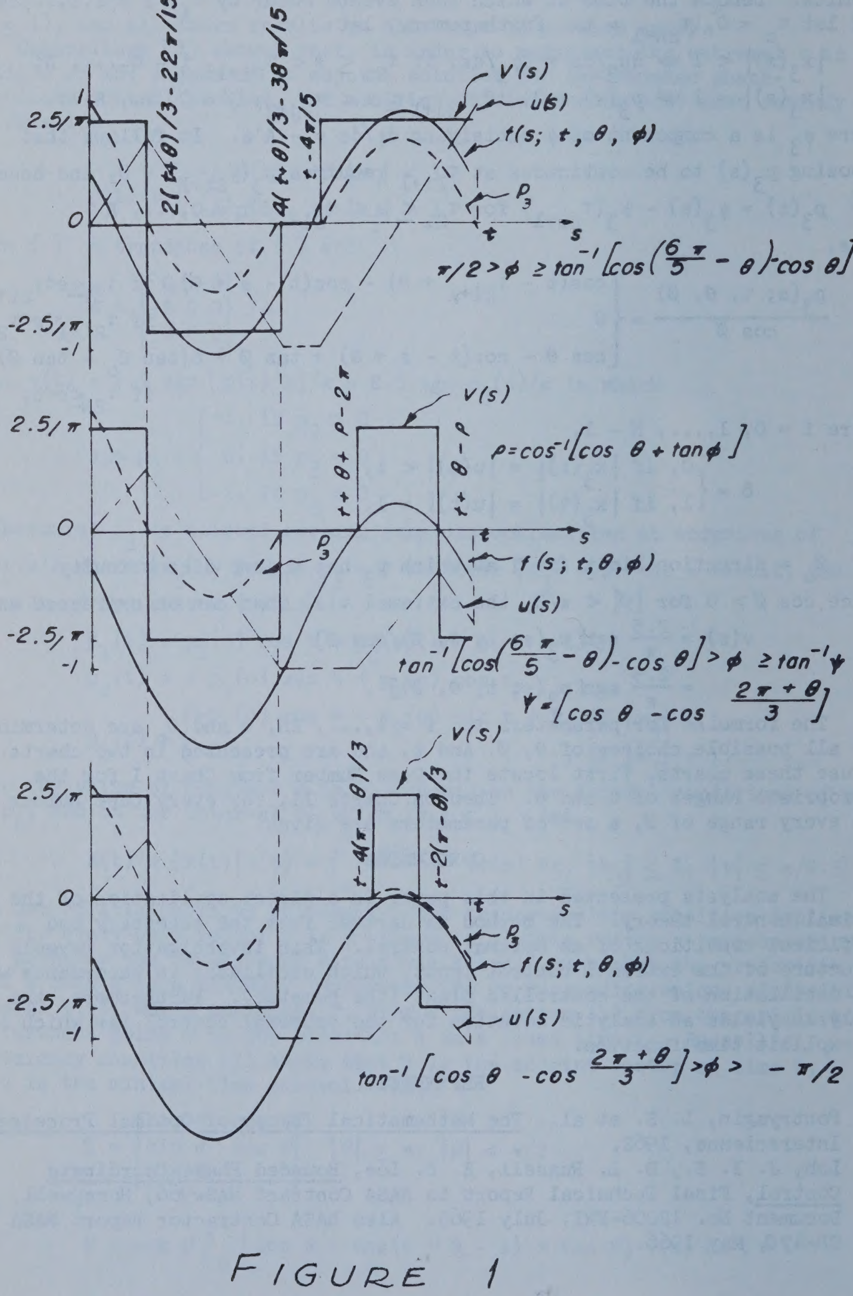

1
2.5/π
0
-2.5/π
-1
2(t+θ)/3 - 22π/15
4(t+θ)/3 - 38π/15
t - 4π/5
V(s)
u(s)
f(s; t, θ, φ)
p_3
t
s
π/2 > φ ≥ tan⁻¹[cos(6π/5 − θ) − cos θ]
2.5/π
0
-2.5/π
-1
t + θ + ρ − 2π
t + θ − ρ
V(s)
ρ = cos⁻¹[cos θ + tan φ]
p_3
t
s
f(s; t, θ, φ)
u(s)
tan⁻¹[cos(6π/5 − θ) − cos θ] > φ ≥ tan⁻¹ψ
ψ = [cos θ − cos (2π + θ)/3]
V(s)
2.5/π
0
-2.5/π
-1
t − 4(π − θ)/3
t − 2(π − θ)/3
t
s
p_3
f(s; t, θ, φ)
u(s)
tan⁻¹[cos θ − cos (2π + θ)/3] > φ > − π/2

FIGURE 1

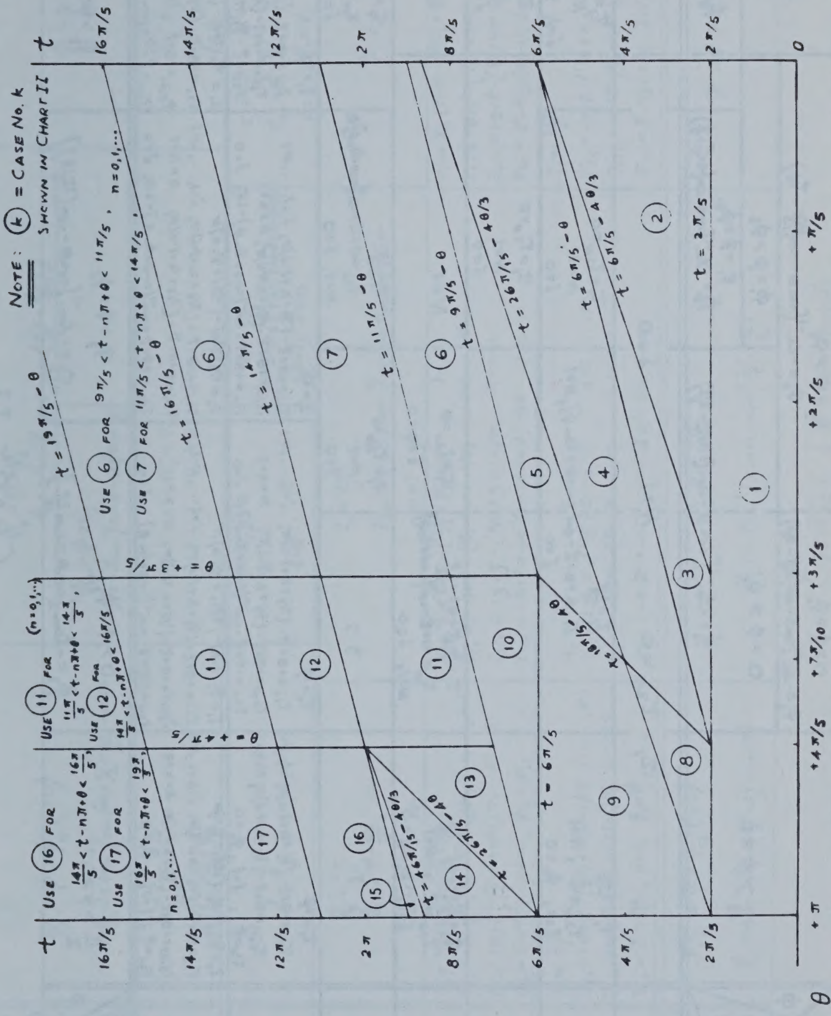

NOTE: (k) = CASE No. k SHOWN IN CHART II
t
θ
0
+ π/5
+ 2π/5
+ 3π/5
+ 7π/10
+ 4π/5
+ π
2π/5
4π/5
6π/5
8π/5
2π
12π/5
14π/5
16π/5
t = 19π/5 − θ
USE (6) FOR 9π/5 < t − nπ + θ < 11π/5 , n = 0, 1, ...
USE (7) FOR 11π/5 < t − nπ + θ < 14π/5 ,
t = 16π/5 − θ
t = 14π/5 − θ
t = 11π/5 − θ
t = 9π/5 − θ
t = 26π/15 − 4θ/3
t = 6π/5 − θ
t = 6π/5 − 4θ/3
t = 2π/5
θ = + 3π/5
θ = + 4π/5
t = 6π/5
t = 18π/5 − 4θ
t = 26π/5 − 4θ
t = 46π/15 − 4θ/3
USE (11) FOR 11π/5 < t − nπ + θ < 14π/5 , (n = 0, 1, ...)
USE (12) FOR 14π/5 < t − nπ + θ < 16π/5
USE (16) FOR 14π/5 < t − nπ + θ < 16π/5 ,
USE (17) FOR 16π/5 < t − nπ + θ < 19π/5 ,
n = 0, 1, ...
1
2
3
4
5
6
6
7
8
9
10
11
11
12
13
14
15
16
17

CHART I

$D=\frac{2.5}{\pi}$ ϕ τ_i's, δ, ϕ_0 CASE	$\pi/2>\phi>0$	$0>\phi>\phi_2$ $\phi_2=-\tan^{-1}[\cos\theta-\cos(\frac{t+\theta}{3}-\frac{6\pi}{5})$	$\phi_2>\phi>\phi_6$ $\phi_6=-\tan^{-1}[\cos\theta-\cos(\frac{11\pi}{10}-\frac{t}{4})$			$\phi_6>\phi>-\pi/2$
		$0>\phi>\phi_1$ $\phi_1=-\tan^{-1}[\cos\theta-\cos(+\theta+t-\frac{2\pi}{5})$		$\phi_1>\phi>\phi_6$		
				$\phi_1>\phi>\phi_4$ $\phi_4=-\tan^{-1}[\cos\theta-\cos(\theta+\frac{\pi}{5}+\frac{t}{2})$	$\phi_4>\phi>-\pi/2$	
1	$\tau_1=\tau_{2N}=0$, $N=1$, $\delta=0$					
2	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=t$, $N=1$ $\delta=1$, $\phi_0=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=+\theta+t-\cos^{-1}[\cos\theta+\tan\phi]$, $N=1$ $\delta=0$		$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=0$	$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=1$, $\phi_0=\phi_4$	
3	))	))		$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=0$		$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=1$, $\phi_0=\phi_6$
4	$\tau_1=\frac{4}{3}(t+\theta)-\frac{6\pi}{5}$ $\tau_{2N}=t$, $N=1$ $\delta=1$, $\phi_0=0$	$\tau_1=4/3(t+\theta)-\frac{6\pi}{5}$ $\tau_{2N}=+\theta+t-\cos^{-1}[\cos\theta+\tan\phi]$ $N=1$, $\delta=0$	$\tau_1=\tau_{2N}=0$, $\delta=0$	$N=1$		))
5	))	))	$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=\theta+t+\cos^{-1}[\cos\theta+\tan\phi]2\pi$ $N=1$, $\delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=t-\frac{4}{3}(\pi-\theta)$, $N=1$ $\delta=1$, $\phi_0=\phi_5$	

6 $n=0,1,\ldots.$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=\theta+t-(3/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N}=t$, $\delta=1$, $\phi_0=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N}=+\theta+t-\cos^{-1}[\cos\theta+\tan\phi]$, $\delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+1$ $\tau_{2N}=+\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $\delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+1$ $\tau_{2N}=t-4/3(\pi-\theta)$, $\delta=1$, $\phi_0=\phi_5$
7 $n=0,1,\ldots.$	$\tau_1=4/3(t+\theta)-(38/15+4/3n)\pi$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N}=t$, $\delta=1$, $\phi_0=0$	$\tau_1=4/3(t+\theta)-(38/15+4/3n)\pi$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N}=+\theta+t-\cos^{-1}[\cos\theta+\tan\phi]$, $\delta=0$	$\tau_1=4/3(t+\theta)-(38/15+4/3n)\pi$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+1$ $\tau_{2N}=+\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $\delta=0$	$\tau_1=4/3(t+\theta)-(38/15+4/3n)\pi$ $\tau_{2i}=+\theta+t-(7/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-(3/5+n+1-i)\pi$, $N=n+1$ $\tau_{2N}=t-4/3(\pi-\theta)$, $\delta=1$, $\phi_0=\phi_5$
CASE τ_i's, δ, ϕ_0 ϕ $D=\frac{2.5}{\pi}$	$\frac{\pi}{2}>\phi>0$	$0>\phi>\phi_3$ $\phi_3=-\tan^{-1}[\cos\theta-\cos\frac{3\pi}{5}]$	$\phi_3>\phi>\phi_5$ $\phi_5=-\tan^{-1}[\cos\theta-\cos(\frac{2\pi+\theta}{3})]$	$\phi_5>\phi>-\frac{\pi}{2}$

CHART II

$D=\frac{2.5}{\pi}$ ϕ / τ_i's, δ, ϕ_0 / CASE	$\pi/2 > \phi > \phi_4\ (>0)$ $\phi_4 = +\tan^{-1}\left[\cos\left(+\theta+\frac{t}{2}-\frac{9\pi}{5}\right)-\cos\theta\right]$	$\phi_4 > \phi > \phi_6$ $\phi_6 = -\tan^{-1}\left[\cos\theta-\cos\left(\frac{11\pi}{10}-\frac{t}{4}\right)\right]$		$\phi_6 > \phi > -\pi/2$
		$\phi_4 > \phi > \phi_1$ $\phi_1 = -\tan^{-1}\left[\cos\theta-\cos\left(+\theta+t-\frac{3\pi}{5}\right)\right.$		
8	$\tau_1 = \tau_{2N} = 0$ $N=1$ $\delta=1,\ \phi_0=\phi_4$	$\tau_1 = \tau_{2N} = 0$ $N=1,\ \delta=0$		$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=1,\ \phi_0=\phi_6$
9	” ”	$\tau_1=\tau_{2N}=0$ $N=1$ $\delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=+\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$ $N=1,\ \delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=t-\frac{4}{3}(\pi-\theta),\ N=1$ $\delta=1,\ \phi_0=\phi_5$

10	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=t-\frac{4\pi}{5},\ N=1$ $\delta=1,\ \phi_0=\phi_{13}$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2N}=+\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$ $N=1,\ \delta=0$	” ”
11 $n=0,1,\ldots\ldots$	$\tau_1=\frac{4}{3}(t+\theta)-\left(\frac{38}{15}+\frac{4}{3}n\right)\pi$ $\tau_{2i}=+\theta+t-\left(\frac{7}{5}+n+1-i\right)\pi,\ i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-\left(\frac{3}{5}+n+1-i\right)\pi,\ N=n+1$ $\tau_{2N}=t-4\pi/5,\ \delta=1,\ \phi_0=\phi_{13}$	$\tau_1=\frac{4}{3}(t+\theta)-\left(\frac{38}{15}+\frac{4}{3}n\right)\pi$ $\tau_{2i}=+\theta+t-\left(\frac{7}{5}+n+1-i\right)\pi,\quad i=1,\ldots N-1$ $\tau_{2i+1}=+\theta+t-\left(\frac{3}{5}+n+1-i\right)\pi,\quad N=n+1$ $\tau_{2N}=+\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi,\ \delta=0$	$\tau_1=\frac{4}{3}(t+\theta)-\left(\frac{38}{15}+\frac{4}{3}n\right)\pi$ $\tau_{2i}=+\theta+t-\left(\frac{7}{5}+n+1-i\right)\pi,\quad i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-\left(\frac{3}{5}+n+1-i\right)\pi,\quad N=n+1$ $\tau_{2N}=t-\frac{4}{3}(\pi-\theta),\ \delta=1,\ \phi_0=\phi_5$
12 $n=0,1,\ldots$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=+\theta+t-\left(\frac{12}{5}+n+1-i\right)\pi,\ i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-\left(\frac{8}{5}+n+1-i\right)\pi,\ N=n+2$ $\tau_{2N}=t-\frac{4\pi}{5},\ \delta=1,\ \phi_0=\phi_{13}$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=+\theta+t-\left(\frac{12}{5}+n+1-i\right)\pi,\quad i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-\left(\frac{8}{5}+n+1-i\right)\pi,\quad N=n+2$ $\tau_{2N}=+\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi,\quad \delta=0$	$\tau_1=\frac{2\pi}{5}$ $\tau_{2i}=+\theta+t-\left(\frac{12}{5}+n+1-i\right)\pi,\quad i=1,\ldots,N-1$ $\tau_{2i+1}=+\theta+t-\left(\frac{8}{5}+n+1-i\right)\pi,\quad N=n+2$ $\tau_{2N}=t-\frac{4}{3}(\pi+\theta),\ \delta=1,\ \phi_0=\phi_5$
CASE / τ_i's, δ, ϕ_0 / ϕ $D=\frac{2.5}{\pi}$	$\pi/2 > \phi > \phi_{13}\ (>0)$ $\phi_{13}=+\tan^{-1}\left[\cos\left(\frac{6\pi}{5}-\theta\right)-\cos\theta\right]$	$\phi_{13} > \phi > \phi_5$ $\phi_5=-\tan^{-1}\left[\cos\theta-\cos\left(\frac{2\pi}{3}+\frac{\theta}{3}\right)\right]$	$\phi_5 > \phi > -\pi/2$

CHART II (CONTINUED)

$D=\frac{2.5}{\pi}$ ϕ τ's, δ, ϕ_0 CASE							
	$\pi/2 > \phi > \phi_{13}$ (>0) $\phi_{13} = +\tan^{-1}[\cos(\frac{6\pi}{5}-\theta)-\cos\theta]$			$\phi_{13} > \phi > \phi_5$			$\phi_5 > \phi > -\pi/2$
	$\pi/2 > \phi > \phi_{12}$ (>0) $\phi_{12} = +\tan^{-1}[\cos(\frac{\theta}{4}-\frac{\pi}{10})-\cos\theta]$			$\phi_{12} > \phi > \phi_{24}$ (>0) $\phi_{24} = +\tan^{-1}[\cos\theta\tan(\frac{t+\theta}{2}-\frac{4\pi}{15})]$		$\phi_{24} > \phi > \phi_5$ $\phi_5 = -\tan^{-1}[\cos\theta-\cos(\frac{2\pi}{3}+\frac{\theta}{3})]$	
	$\pi/2 > \phi > \phi_{23}$ (≥0) $\phi_{23} = +\tan^{-1}[\cos(\frac{2\pi}{3}-\frac{\theta}{3})-\cos\theta]$	$\phi_{23} > \phi > \phi_1$ (>0) $\phi_1 = +\tan^{-1}[\cos\theta\tan(\theta+t-\frac{2\pi}{5})]$			$\phi_1 > \phi > \phi_{24}$		
13	$\tau_1 = \frac{4}{3}(+\theta+t)-\frac{38\pi}{15}$ $\tau_{2N} = t-\frac{4\pi}{5}$, $N=1$ $\delta=1$, $\phi_0=\phi_{13}$			$\tau_1 = 4/3(+\theta+t)-\frac{38\pi}{15}$ $\tau_{2N} = +\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $N=1$ $\delta=0$			$\tau_1 = 4/3(+\theta+t)-\frac{38\pi}{15}$ $\tau_{2N} = t-4/3(\pi-\theta)$, $N=1$ $\delta=1$, $\phi_0=\phi_5$
14	$\tau_1 = \tau_{2N} = 0$ $N=1$ $\delta=1$, $\phi_0=\phi_{12}$		$\tau_1 = \tau_{2N} = 0$, $N=1$ $\delta=0$			$\tau_1 = 4/3(+\theta+t)-\frac{38\pi}{15}$ $\tau_{2N} = +\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$ $N=1$, $\delta=0$	〃
15	$\tau_1 = \frac{2\pi}{5}$ $\tau_{2N} = t-4/3(2\pi-\theta)$ $N=1$, $\delta=1$, $\phi_0=\phi_{23}$	$\tau_1 = \frac{2\pi}{5}$ $\tau_{2N} = \theta+t-\cos^{-1}[\cos\theta+\tan\phi]-2\pi$ $N=1$, $\delta=0$			$\tau_1 = \tau_{2N} = 0$ $N=1$ $\delta=0$	〃	〃

16 $n=0,1,\ldots$	$\tau_1 = \frac{2\pi}{5}$ $i=1,\ldots,N-1$ $\tau_{2i} = \theta+t-(\frac{12}{5}+n+1-i)\pi$ $\tau_{2i+1} = \theta+t-(\frac{8}{5}+n+1-i)\pi$, $N=n+1$ $\tau_{2N} = t-4/3(2\pi-\theta)$, $\delta=1$, $\phi_0=\phi_{23}$	$\tau_1 = \frac{2\pi}{5}$ $\tau_{2i} = \theta+t-(\frac{12}{5}+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = \theta+t-(8/5+n+1-i)\pi$, $N=n+1$ $\tau_{2N} = +\theta+t-\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $\delta=0$	$\tau_1 = \frac{2\pi}{5}$ $\tau_{2i} = +\theta+t-(\frac{12}{5}+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = +\theta+t-(8/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N} = \theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $\delta=0$	$\tau_1 = \frac{2\pi}{5}$ $\tau_{2i} = +\theta+t-(12/5+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = +\theta+t-(8/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N} = t-4/3(\pi-\theta)$, $\delta=1$, $\phi_0=\phi_5$
17 $n=0,1,\ldots$	$\tau_1 = 4/3(+\theta+t)-(\frac{58}{15}+4/3\,n)\pi$ $\tau_{2i} = \theta+t-(\frac{12}{5}+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = \theta+t-(\frac{8}{5}+n+1-i)\pi$, $N=n+1$ $\tau_{2N} = t-4/3(2\pi-\theta)$, $\delta=1$, $\phi_0=\phi_{23}$	$\tau_1 = 4/3(+\theta+t)-(\frac{58}{15}+4/3\,n)\pi$ $\tau_{2i} = +\theta+t-(\frac{12}{5}+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = +\theta+t-(8/5+n+1-i)\pi$, $N=n+1$ $\tau_{2N} = \theta+t-\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $\delta=0$	$\tau_1 = 4/3(+\theta+t)-(\frac{58}{15}+4/3\,n)\pi$ $\tau_{2i} = +\theta+t-(\frac{12}{5}+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = +\theta+t-(\frac{8}{5}+n+1-i)\pi$, $N=n+2$ $\tau_{2N} = +\theta+t+\cos^{-1}[\cos\theta+\tan\phi]-2\pi$, $\delta=0$	$\tau_1 = 4/3(+\theta+t)-(\frac{58}{15}+4/3\,n)\pi$ $\tau_{2i} = +\theta+t-(\frac{12}{5}+n+1-i)\pi$, $i=1,\ldots,N-1$ $\tau_{2i+1} = \theta+t-(8/5+n+1-i)\pi$, $N=n+2$ $\tau_{2N} = t-4/3(\pi-\theta)$, $\delta=1$, $\phi_0=\phi_5$
CASE τ's, δ, ϕ_0 ϕ $D=\frac{2.5}{\pi}$	$\pi/2 > \phi > \phi_{23}$ (>0) $\phi_{23} = +\tan^{-1}[\cos(\frac{2\pi}{3}-\frac{\theta}{3})-\cos\theta]$	$\phi_{23} > \phi > \phi_{25}$ (>0) $\phi_{25} = +\tan^{-1}[\cos\frac{2\pi}{5}-\cos\theta]$	$\phi_{25} > \phi > \phi_5$ $\phi_5 = -\tan^{-1}[\cos\theta-\cos(\frac{2\pi}{3}+\frac{\theta}{3})]$	$\phi_5 > \phi > -\pi/2$

CHART II (CONTINUED)

OPTIMIZATION AND IMPLEMENTATION OF SYSTEM CONTROL LAWS*

M. Schoenberger
Surface Division, Westinghouse Defense and Space Center, Baltimore 3, Md.

ABSTRACT

A method is presented for optimizing the performance of a system without using excessively complex controllers. In order to insure ease of implementation, a feedback arrangement involving time-invariant parameters is introduced as a constraint on the problem. Necessary conditions for optimization with respect to these parameters are obtained from the calculus of variations, and algorithms are presented for obtaining numerical solutions. An application of this procedure to submarine diving control is discussed, and the performances associated with various control structures are compared.

I. INTRODUCTION

Most of the research in optimal control systems is concerned with determining the control $\underline{u}(t)$ in the class of piecewise-continuous vector functions which minimizes a performance index. This control is usually obtained by application of the calculus of variations or the Maximum Principle, each of which provides a set of necessary conditions for optimality in the form of a two-point boundary value problem.

However, there has been little concern with the eventual implementation of the optimal control. Even in those cases in which the two-point boundary value problem can be solved analytically, the control is often difficult to implement [1] . Difficult implementation is, in fact, assured in the majority of cases where the problem must be solved by numerical techniques such as the method of steepest ascent [2] . In these cases, the control is an optimal sequence of vectors $\underline{u}(0)$, $\underline{u}(\Delta T)$, ..., $\underline{u}(k\Delta T)$, ..., $\underline{u}(T)$ which must not only be stored for each set of initial conditions, but must be applied to the system by using some type of a computer.

If a relatively simple control is desired, some practical compromise must be made prior to implementation. Therefore, in this paper, a control structure is specified at the outset, and the performance is optimized subject to this structural constraint.

Sobral [3] and Rohrer and Sobral [4] have solved special cases of the implementation problem involving linear systems and singular quadratic performance indices. Rekasius [5] and Bass and Webber [6] have determined

---

* The results reported here are based upon part of a doctoral dissertation in Electrical Engineering submitted to the University of Illinois, Urbana, January 1966. This work was supported in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract No. DA 28 043 AMC 00073(E) and in part by the Air Force Office of Scientific Research under Grant AF AFOSR 931.65.

procedures for obtaining sub-optimal controls that are easier to implement than the optimal. However, these controls only approximate the optimal, and using the same feedback arrangement, better performance can be obtained.

II. PROBLEM FORMULATION

The system differential equations to be considered are

$$\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{u}, t) \tag{1}$$

where $\underline{x}$ is an n-vector of the states, $\underline{u}$ is an m-vector of the control variables, and t is time. The initial conditions are known.

$$\underline{x}(t_0) = \underline{x}_0. \tag{2}$$

The performance index is

$$J = \phi\left[\underline{x}(T), T\right] + \int_{t_0}^{T} L(\underline{x}, \underline{u}, t)dt \tag{3}$$

where the initial time t_0 and the terminal time T are fixed. Furthermore, the feedback arangement is specified to be

$$\underline{u} = \underline{g}(\underline{x}, \underline{c}, t) \tag{4}$$

$$\dot{\underline{c}} = \underline{0}, \tag{5}$$

where $\underline{g}$ is an algebraic relationship among $\underline{x}$, $\underline{c}$, and t, and $\underline{c}$ is an r-vector containing the parameters of the control structure.

The problem to be considered is: Choose the control $\underline{u}$, from the set of admissible controls, which minimizes J subject to the n differential constraints (1). The set of admissible controls consists of those piecewise-continuous vector functions that satisfy (4) and (5).

Substituting (4) into $\underline{f}(\underline{x}, \underline{u}, t)$ and $L(\underline{x}, \underline{u}, t)$ results in $\underline{f}'(\underline{x}, \underline{c}, t)$ and $L'(\underline{x}, \underline{c}, t)$, respectively; the system equations

$$\dot{\underline{x}} = \underline{f}'(\underline{x}, \underline{c}, t) \tag{6a}$$

$$\dot{\underline{c}} = \underline{0}, \tag{6b}$$

and the performance index

$$J' = \phi\left[\underline{x}(T), T\right] + \int_{t_0}^{T} L'(\underline{x}, \underline{c}, t)dt. \tag{7}$$

Necessary conditions for an extremum of J' subject to the n+r differential constraints of (6) are the equations (6), themselves, and the following Euler equations and transversality conditions [7]

$$\dot{\underline{\lambda}} = -\nabla_{\underline{x}}H, \quad \underline{\lambda}(T) = \nabla_{\underline{x}(T)}\phi\left[\underline{x}(T), T\right] \tag{8}$$

$$\dot{\underline{\gamma}} = \nabla_{\underline{c}}H, \quad \underline{\gamma}(t_0) = \underline{\gamma}(T) = \underline{0} \tag{9}$$

where the Hamiltonian H is

$$H \triangleq L'(\underline{x}, \underline{c}, t) + \underline{\lambda}^T \underline{f}'(\underline{x}, \underline{c}, t) \tag{10}$$

and $\nabla_{\underline{a}} H$ is a p-vector whose ith element is $\frac{\partial H}{\partial a_i}$ when $\underline{a}$ is a p-vector. If $\underline{x}(t_0)$ is not specified, the boundary condition $\underline{\lambda}(t_o) = \underline{0}$ replaces (2).

Equations (2), (6), (8) and (9) comprise a set of necessary conditions for optimization consisting of 2(n+r) differential equations and 2(n+r) boundary values, of which n+r are specified at $t = t_o$ and n+r at t = T. The difficulty in solving this type of two-point boundary value problem provides the incentive to modify the formulation.

Since (9) implies

$$\int_{t_o}^{T} \nabla_{\underline{c}} H \, dt = \underline{0} \tag{11}$$

equations (2), (6), (8) and (11) constitute a set of necessary conditions for minimizing J'. The solution to this set of equations is obtained in a manner similar to the method of steepest ascent [2]. Assuming a value for $\underline{c}$, $\underline{x}(t)$ ($t_o \leq t \leq T$) is obtained by integrating (6) in the forward direction. $\underline{\lambda}(T)$ can then be calculated from (8) and $\underline{\lambda}(t)$ ($t_o \leq t \leq T$) obtained by integrating (8) in the reverse direction. Since this process satisfies all of the conditions but (11), its success depends upon choosing $\underline{c}$ in such a fashion that the left side of (11) becomes successively closer to $\underline{0}$. In the next two sections, this optimization procedure is applied to two classes of problems, and algorithms are developed for the satisfaction of (11).

III. LINEAR SYSTEMS WITH OPTIMAL LINEAR CONTROL

Consider the class of problems in which the plant is linear in the control and the performance index is quadratic in the control; i.e.,

$$\dot{\underline{x}} = \underline{f}(\underline{x}) + B(\underline{x}) \underline{u} \tag{12}$$

with the initial conditions specified as in (2) where $\underline{x}$ is an n-vector, $\underline{u}$ is an m-vector and $B(\underline{x})$ is nxm. The performance index is

$$J = \underline{x}^T(T) P \underline{x}(T) + \int_0^T \left[L(\underline{x}) + \underline{u}^T R \underline{u} \right] dt, \tag{13}$$

where $L(\underline{x})$ is positive semi-definite, R is a diagonal positive-definite matrix, and P is a symmetric positive-definite matrix. Since it is easy to implement, the control is constrained to be

$$\underline{u} = C\underline{x}, \quad \dot{C} = 0 \tag{14}$$

where C is an mxn matrix. (Actually, a diagonal R is not essential to the problem; it merely simplifies the equations.)

If $\underline{c}_i$ represents the vector formed from the ith row of C and $\underline{b}_i$ the vector from the ith column of $B(\underline{x})$, the Hamiltonian for this system is

$$H = L(\underline{x}) + \underline{\lambda}^T\underline{f}(\underline{x}) + \sum_{i=1}^{m} \left[r_{ii}(\underline{c}_i^T\underline{x}\underline{x}^T\underline{c}_i) + \underline{\lambda}^T\underline{b}_i(\underline{x})\underline{c}_i^T\underline{x} \right] \tag{15}$$

From Section II, necessary conditions for J' to be a minimum are equations (2),

$$\dot{\underline{x}} = \underline{f}(\underline{x}) + B(\underline{x})\, C\underline{x}, \quad \dot{C} = 0 \tag{16}$$

$$\dot{\underline{\lambda}} = -\nabla_{\underline{x}}H, \quad \underline{\lambda}(T) = 2P\underline{x}(T) \tag{17}$$

and

$$r_{ii}P_1\underline{c}_i + 1/2\,\underline{\beta}_i = \underline{0}, \quad i = 1, 2, \ldots, m \tag{18}$$

where

$$P_1 \triangleq \int_0^T \underline{x}(t)\underline{x}^T(t)dt, \quad \underline{\beta}_i \triangleq \int_0^T \underline{x}(t)\underline{\lambda}^T(t)\underline{b}_i(\underline{x})dt. \tag{19}$$

An iterative procedure for solving these equations involves rewriting (18) as

$$\underline{c}_i = -\frac{1}{2r_{ii}} P_1^{-1}\underline{\beta}_i \triangleq \underline{Z}_i(C) \tag{20}$$

Denoting the matricies formed during the kth iteration by the superscript k and assuming a value for $C^{(1)}$, the iterative procedure is:

(1) Let $\underline{x}^{(k)}(0) = \underline{x}_o$, integrate (16) in the forward direction and store $\underline{x}^{(k)}(t)$, $t_o \leq t \leq T$. $P_1^{(k)}$ can be formed either by simultaneously integrating (19) or, in the linear case, analytically [8].

(2) From (17), compute $\underline{\lambda}^{(k)}(T)$ and integrate (17) backward from t = T to 0. Also, determine $\underline{\beta}_i^{(k)}$ by simultaneously integrating (19).

(3) Form

$$\underline{y}_i^{(k)} = r_{ii}P_1^{(k)}\underline{c}_i^{(k)} + 1/2\underline{\beta}_i^{(k)}, \quad i = 1, 2, \ldots, m, \tag{21}$$

and terminate the procedure if $\sum_{i=1}^{m} \|\underline{y}_i^{(k)}\|$ is less than some preselected value. Otherwise, let $\underline{c}_i^{(k+1)} = \underline{Z}_i(C^{(k)})$, i = 1, 2, ..., m, return to step 1, and repeat until the termination criterion is satisfied.

The process described above is the method of successive substitutions [9] which converges for $\theta_{ij} \leq 1$, i = 1, 2, ..., m, j = 1, 2, ..., n, in the neighborhood of $C = C^{\circ}$.

where

$$\theta_{ij} \triangleq \sum_{k=1}^{n} \left| \frac{(\underline{Z}_i)_j}{c_{ik}} \right|$$

If $\theta_{ij} \ll 1$, i = 1, 2, ..., m; j = 1, 2, ..., n, the convergence is rapid.

Since the region in which the convergence condition is satisfied is seldom a priori known, $C^{(i)}$ cannot always be chosen within this region. Therefore, occasionally $\sum_{i=1}^{m} \left\| \underline{y}_i^{(k+1)} \right\| \geq \sum_{i=1}^{m} \left\| \underline{y}_i^{(k)} \right\|$, and it is necessary to repeat the iteration after choosing $C^{(k+1)}$ by another method. Despite occasional unsuccessful iterations, this procedure has converged quite rapidly when applied to problems involving linear systems with quadratic and quartic performance indices.

IV. REGULATOR PROBLEM

One of the more popular problems in optimal control is the linear regulator problem. In this problem, the plant is described by

$$\dot{\underline{x}} = A\underline{x} + B\underline{u} \tag{22}$$

where A and B are constant nxn and nxm matrices, respectively, and the initial conditions are specified as in (2). The performance index is

$$J = \underline{x}^T(T)P\underline{x}(T) + \int_0^T \left[\underline{x}^T Q\underline{x} + \underline{u}^T R\underline{u}\right] dt, \tag{23}$$

where P and Q are symmetric and positive semi-definite and R is symmetric and positive definite.

The optimal control has been shown to be linear time-varying feedback [10], $\underline{u}^o = M(t)\underline{x}$. However, since the gains of the feedback amplifiers are unrealizable if P is sufficiently large and time-varying feedback is difficult to implement, Rekasius [1] obtains an alternate solution to this problem in the form of time-invariant feedback plus open loop control, $u^o = \underline{h}^T\underline{x} + v(t)$.

This solution has the same implementation disadvantage as most open-loop controls. It must be determined numerically, stored and then applied to the system in the required sequence. Certainly, the following arrangement is a considerably simpler controller.

$$\underline{u} = C\,\underline{x} + \underline{d}, \quad \dot{C} = 0, \quad \dot{\underline{d}} = \underline{0} \tag{24}$$

From Section II, necessary conditions for J' to be a minimum are (2), (8), and

$$\dot{\underline{x}} = (A + BC)\underline{x} + B\underline{d}, \quad \dot{C} = 0, \quad \dot{\underline{d}} = \underline{0} \tag{25}$$

$$\left[\begin{array}{c|c} r_{ii}P_1 & \underline{p}_2 \\ \hline \underline{p}^T_2 & T \end{array}\right] \left[\begin{array}{c} \underline{c}_i \\ \hline d_i \end{array}\right] + 1/2 \left[\begin{array}{c} \underline{\beta}_i \\ \hline \alpha_i \end{array}\right] = \underline{0} \tag{26}$$

for $i = 1, 2, \ldots, m$

where P_1 and $\underline{\beta}_i$ are defined by (19) and

$$\underline{p}_2 \triangleq \int_0^{60} \underline{x}(t)dt, \quad \alpha_i \triangleq \frac{1}{r_{ii}} \int_0^{60} \underline{\lambda}^T \underline{b}_i dt \tag{27}$$

A seemingly good iterative procedure for solving these equations is the method of successive substitutions presented in the previous section in which

$$\begin{bmatrix} c_i^{(k+1)} \\ d_i^{(k+1)} \end{bmatrix} = -1/2 \begin{bmatrix} r_{ii}P_1^{(k)} & p_2^{(k)} \\ p_2^{T(k)} & T \end{bmatrix}^{-1} \begin{bmatrix} \beta_i^{(k)} \\ \alpha_i^{(k)} \end{bmatrix} \qquad (28)$$

However, as mentioned earlier, this type of procedure converges rapidly only when the partial derivatives with respect to c_{ij} and d_i of P_1, $\underline{p}_2$, $\underline{\beta}_i$, and α_i are much less than one in absolute value. Since $\underline{p}_2$ and $\underline{\alpha}_i$ have a linear dependence on $\underline{d}$, and P_1 and $\underline{\beta}_i$ have both a linear and quadratic dependence on $\underline{d}$, this may not be the case. For this reason and to reduce computer time per iteration, it is useful to derive expressions for these matrices which depend explicitly on C and $\underline{d}$ [8] .

The method of successive substitutions presented earlier can then be utilized after these formulae are substituted into (26). $\underline{c}_i^{(k+1)}$ and $d_i^{(k+1)}$ are chosen such that the left side of (26) is equal to $\underline{0}$. This requires the solution to a set of nonlinear algebraic equations which is easily accomplished numerically.

A Submarine-Diving Example

In this section, the approach developed previously is applied to submarine diving by use of stern plane control. Consider a model in which x_1 is the pitch angle, x_2 is the pitch rate, x_3 is the angle of attack and u is the stern plane angle of deflection. x_4 = (depth-desired depth)/100. Assuming a constant forward speed of 25 ft/sec. the linearized equations of motion in the vertical plane are [8, 11] .

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1.0 & 0 & 0 \\ -.0071 & -.111 & .12 & 0 \\ 0 & .07 & -.30 & 0 \\ -.25 & 0 & .25 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ -.095 \\ .072 \\ 0 \end{bmatrix} u \qquad (29)$$

or, in vector notation, $\dot{\underline{x}} = A\underline{x}+\underline{b}u$.

The submarine is to dive in 60 seconds from the water's surface to a depth of 200 ft. and remain there. The trajectory must be as low as is consistent with a soft magnitude constraint of 1.5° on the stern plane angle.

$$\underline{x}_0 = [0, 0, 0, -2.]^T \qquad (30)$$

Interpreting the "lowest trajectory" to mean minimum area under the depth error curve,

$$J = \underline{x}^T(60)P\underline{x}(60) + \int_0^{60} \left[\underline{x}^T Q\underline{x} + u^2\right] dt \qquad (31)$$

where Q is chosen such that $x_4^{\,2}$ is weighted heavily and the other state variables and u are constrained in magnitude.

$$Q = \begin{bmatrix} 10^{-3} & 0 & 0 & 0 \\ 0 & 10^{-3} & 0 & 0 \\ 0 & 0 & 10^{-3} & 0 \\ 0 & 0 & 0 & 10^{-4} \end{bmatrix} \qquad (32)$$

P is a positive definite diagonal matrix whose elements are large enough to force $\underline{x}(60)$ to be very small; thus, causing the submarine to remain at a depth of 200 ft.

The optimal control, $u^o(t)$ is [1]

$$u^o(t) = \underline{c}_o^T \underline{x} + V(t) \qquad (33)$$

where $\underline{c}_o = \left[.0296,\ .226,\ .412,\ -.918 \times 10^{-2}\right]^T$ and V(t) is plotted in Figure 1. The difficulty in implementing V(t) makes it worthwhile to compare its performance to that resulting from simpler controls.

If the system is not completely observable or if some of the state variables are not readily available, it may be desirable to feedback only the depth error; i.e., J is to be minimized subject to (29) and

$$u_1 = K x_4, \quad \dot{K} = 0. \qquad (34)$$

Applying the procedure described earlier results in 2n+1 differential equations given by equations (16) and (8a), 2n boundary values given by (2) and (8b), and

$$2K \int_0^{60} x_4^2 \, dt + \int_0^{60} \underline{\lambda}^T \underline{b} x_4 \, dt = 0. \qquad (35)$$

Therefore, the method of successive substitutions described in Section III can once more be applied using the algorithm

$$K^{(i+1)} = -\frac{1}{2} \int_0^{60} (\underline{\lambda}^{(i)})^T \underline{b} x_4^{(i)} dt \Big/ \int_0^{60} (x_4^{(i)})^2 dt$$

The feedback parameter resulting from this optimization is K = -.0116, and the associated diving pattern and control are shown in Figures 2 and 3, respectively.

Another control that can be easily implemented is

$$u_2 = \underline{c}_2^T \underline{x}, \quad \dot{\underline{c}}_2 = \underline{0}$$

The necessary conditions for optimality in this case are given by (17), (18) and (19). The iterative procedure discussed following these equations results in the optimal parameters $\underline{c}_2 = (-.014, -.020, -1.230, -.013)^T$. The diving pattern and control associated with u_2 are shown in Figures 2 and 3, respectively. Since by far the most significant contribution to u_2 is the depth error, the similarity in performance of these controls is to be expected. An additional degree of complexity is provided by

$$u_3 = \underline{c}_3^T \underline{x} + d, \quad \dot{\underline{c}}_3 = \underline{0}, \ \dot{d} = 0.$$

Utilizing the procedure of the preceding section results in the optimal parameters

$$\underline{c}_3 = \left[.0127, 1.66, -24.5, -.0412\right]^T, \quad d = .0417.$$

An examination of Figures 2 and 3 provides an overall contrast of the system's response to each of the controls. The weighting of terms in the integrand of the performance index has resulted in each of the controls being of the same order of magnitude during most of the time interval.

Figure 2 graphically illustrates what one would qualitatively expect; there is a trade-off between performance, as measured by the diving pattern, and ease of implementation. As the constraints on the control are relaxed, better performance is obtained. The best performance is obtained from u^o, with u_3, u_2 and u_1 being less optimal in that order.

However, it must be noted that despite large differences in ease of implementation, all of the trajectories are quite similar. The designer must make a qualitative decision as to whether a small gain in performance is worth the implementation of a more complex controller. Certainly u_2 would not be a very popular control since a negligible gain in performance over u_1 requires feedback of three additional state variables.

V. CONCLUSIONS

Necessary conditions for optimal performance of a system in which the control structure is specified have been derived using the calculus of variations. The difficulties involved in obtaining an analytic solution to the resultant two-point boundary value problem motivated its reformulation into the well known method of successive substitutions.

This method of optimization was programmed on a computer, and the practical problem of optimal submarine diving was investigated in considerable detail. The four types of control that were compared were optimal control, linear feedback plus a constant, linear feedback, and feedback of only the depth error. As was expected, better performance resulted as the restrictions on the control were relaxed, but in some cases the performance gain was quite small. In general, the designer must make a qualitative decision based on the merits of the specific problem as to whether the gain in performance of the system is worth the increased complexity of the controller.

A procedure has also been developed for optimizing systems with fixed control structure when a probability distribution of initial conditions is specified [8]. Since these results are primarily useful when the terminal time is infinite, the present paper complements them in the case of known initial conditions and finite terminal time. Another class of problems which should be investigated is that in which the optimal control is bang-bang. For the minimal-time or minimal-fuel problem, $u^o = \text{sgn}\left[g(\underline{x})\right]$ where $g(\underline{x})$ is a nonlinear function of $\underline{x}$. Due to the difficulty in implementing $g(\underline{x})$, it would be worthwhile to compare the performance resulting from a control of the form $u = \text{sgn}\,(\underline{c}^T\underline{x})$.

BIBLIOGRAPHY

1. Rekasius, Z. V., "An Alternate Approach to the Fixed Terminal Point Regulator Problem," IEEE Trans. on Automatic Control, Vol. AC-9, No. 3, pp. 290-292, July, 1964.

2. Bryson, A. E. and Denham, W. F., "A Steepest-Ascent Method for Solving Optimum Programming Problems," Trans. of the ASME Journal Of Applied Mechanics, Vol. 29, pp. 247-257, June, 1962.

3. Sobral, M., "Linear Control Laws for Singular Linear Systems," Report R-188, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, January, 1964.

4. Rohrer, R. A. and Sobral, M., "Optimal Linear Switching for Singular Linear Systems," Report R-196, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, March, 1964.

5. Rekasius, Z. V., "Suboptimal Design of Intentionally Nonlinear Controllers," IEEE Trans. on Automatic Control, Vol. AC-9, No. 4, pp. 380-386, October, 1964.

6. Bass, R. W. and Webber, R. F., "On the Sub-Optimal Control of Autonomous Linear Systems," Hughes Aircraft Co. Tech. Rept., March, 1965.

7. Elsgolc, L. E., Calculus of Variations, Addison Wesley Publishing Company, Reading, Massachusetts, 1962.

8. Schoenberger, M., "System Optimization with Fixed Control Structure," Report R-278, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, January, 1966.

9. Scarborough, J. B., Numerical Mathematical Analysis, The Johns Hopkins Press, Baltimore, Maryland, pp. 209-210, 1950.

10. Kalman, R. E., "The Theory of Optimal Control and the Calculus of Variations," Mathematical Optimization Techniques, ed. R. Bellman, University of California Press, Berkeley, California, 1963.

11. Hoffman, C. F., "Control of Submarine Maneuvers," Discrete Systems Concepts Project, Monograph No. 3, Illinois Institute of Technology, pp. 1-24, June, 1965.

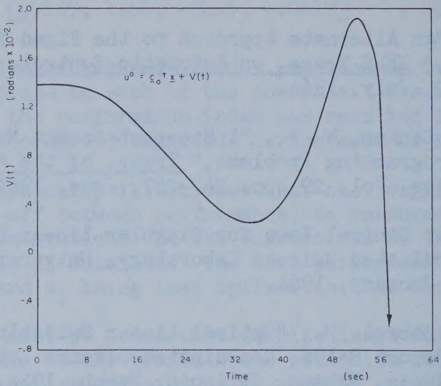


Figure 1. Open-loop term of optimal control.

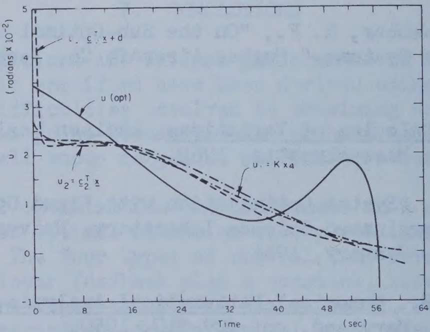


Figure 2. Control (stern plane angle of deflection) vs. time for each of the control configurations

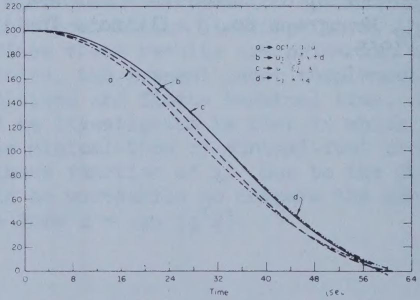


Figure 3. Depth error vs. time for each of the control configurations

ON THE QUADRATIC CONTROL PROBLEM WITH AN ENERGY CONSTRAINT

Gerald Cook, Frank J. Seiler Research Laboratory,
USAF Academy, Colorado
James E. Funk, Graduate student, Astronautics
Department, University of Michigan, Ann Arbor, Mich

ABSTRACT

The quadratic control regulator problem with an energy constraint has been investigated from the viewpoint of having limited control energy available or as an indirect method of limiting the maximum control amplitude. The solution of this problem requires iteration on a parameter λ_{n+1} and therefore is not as straightforward to apply as the well known Kalman result where no constraints are imposed. However, for the systems studied it was found that λ_{n+1} had a regular behavior as a function of initial state and that this behavior could be approximated by empirical equations. Therefore, the control law could be approximately known for all initial conditions without iterating. Also the numerical results indicated that constraining control energy did tend to constrain maximum control amplitude.

I. INTRODUCTION

A well known result in the field of optimal control is R.E. Kalman's [1] solution to the quadratic control problem. The most general result is the following:

Given the linear system with state $\underline{x}(t)$ and control $\underline{u}(t)$ and described by the vector differential equation:

$$\dot{\underline{x}}(t) = \underline{A}(t)\ \underline{x}(t) + \underline{B}(t)\ \underline{u}(t); \quad \underline{x}(t_o) = \underline{\xi} \tag{1}$$

the control that minimizes the functional:

$$J = \frac{1}{2}\ \underline{e}(t)'\ \underline{F}\ \underline{e}(t) + \frac{1}{2}\int_{t_o}^{T} [\underline{e}(t)'\ \underline{Q}(t)\ \underline{e}(t) + \underline{u}(t)'\ \underline{R}(t)\ \underline{u}(t)]dt \tag{2}$$

where:

$$\underline{e}(t) = \underline{Z}(t) - \underline{x}(t)$$

is the error vector between the state $\underline{x}(t)$ and the desired response $\underline{Z}(t)$; $\underline{F}$ and $\underline{Q}(t)$ are positive semidefinite matrices; $\underline{R}(t)$ is a positive definite matrix; t_o is the initial time, and T is the fixed terminal time, is given by:

$$\underline{u}(t) = \underline{R}^{-1}(t)\ \underline{B}'(t)[\underline{g}(t) - \underline{K}(t)\ \underline{x}(t)] \tag{3}$$

where:

the matrix $\underline{K}(t)$ is the solution of the matrix Riccati differential equation

$$\dot{\underline{K}}(t) = -\underline{K}(t)\ \underline{A}(t) - \underline{A}'(t)\ \underline{K}(t) + \underline{K}(t)\ \underline{B}(t)\ \underline{R}^{-1}(t)\ \underline{B}'(t)\ \underline{K}(t) - \underline{Q}(t) \tag{4}$$

with the boundary condition

$$\underline{K}(T) = \underline{F}$$

and the vector $\underline{g}(t)$ is the solution of the linear vector differential equation

$$\dot{\underline{g}}(t) = -[\underline{A}(t) - \underline{B}(t)\,\underline{R}^{-1}(t)\,\underline{B}'(t)\,\underline{K}(t)]'\,\underline{g}(t) - \underline{Q}(t)\,\underline{Z}(t) \qquad (5)$$

with the boundary condition

$$\underline{g}(T) = \underline{F}\,\underline{Z}\,(T)\ .$$

It is clear that if the desired response, $\underline{Z}(t)$, is time varying, then the present value of the optimum control depends on future values of the desired response $\underline{Z}(t)$. The simplest case occurs when $\underline{Z}$ is zero. Then

$$\underline{g}(t) = \underline{0} \qquad (6)$$

for all time and one has the much simpler control law:

$$\underline{u}(t) = -\underline{R}^{-1}(t)\,\underline{B}'(t)\,\underline{K}(t)\,\underline{x}(t)\ . \qquad (7)$$

One drawback to this development is that no constraints on the control variable are employed. The present paper investigates a related but different problem. The cost criterion is taken to be:

$$J = \frac{1}{2}\,\underline{e}(T)'\,\underline{F}\,\underline{e}(T) + \frac{1}{2}\int_{t_o}^{T} \underline{e}(t)'\underline{Q}(t)\,\underline{e}(t)\,dt \qquad (8)$$

with the constraint

$$G = \frac{1}{2}\int_{t_o}^{T} \underline{u}'(t)\,\underline{R}(t)\,\underline{u}(t)\,dt \le E\ . \qquad (9)$$

Here one takes into account the fact that there may be a limited amount of control energy available for the process. It is also known that limiting the integral square value of a function tends to limit the maximum absolute value of the function. Thus one may use an energy constraint as an indirect means of constraining $|u|_{max}$ if such a constraint is required.

Newton [2] treated a special case of this problem requiring that $\underline{A}$ and $\underline{B}$ be time invariant, $\underline{F} = \underline{0}$, $\underline{Q}$ be time invariant and that the control interval be infinite. With these assumptions the Weiner-Hopf equation was desired and solutions were obtained. However, the usual method of solving the Weiner-Hopf equation, spectral factorization, cannot be used if the time interval is finite, if the system is time varying, or if $\underline{Q}$ is time varying. For these reasons a different approach is taken, making use of the maximum principle of Pontryagin.

II. THEORETICAL DEVELOPMENT

Only the special problem, $\underline{Z} = \underline{0}$, will be considered here. Extension to the more general problem is straightforward in theory but would result in more complex problems for numerical evaluation.

The cost function now becomes

$$J = \frac{1}{2}\,\underline{x}(T)'\,\underline{F}\,\underline{x}(t) + \frac{1}{2}\int_{t_o}^{T} \underline{x}(t)'\,\underline{Q}(t)\,\underline{x}(t)\,dt \,. \tag{10}$$

We wish to minimize J while satisfying the constraint

$$G = \frac{1}{2}\int_{t_o}^{T} \underline{u}'(t)\,\underline{R}(t)\,\underline{u}(t)\,dt \le E \tag{11}$$

As before, the differential equations governing the system are:

$$\dot{\underline{x}} = \underline{A}(t)\,\underline{x} + \underline{B}(t)\,\underline{u} \tag{12}$$

where: $\underline{F}$ and $\underline{Q}(t)$, n×n, are positive semidefinite, R(t), m×m, is positive definite

$\underline{x}$ is n×1, $\underline{A}(t)$ is n×n, $\underline{u}$ is m×1 and $\underline{B}(t)$ is n×m.

By taking

$$\dot{x}_{n+1} = \frac{1}{2}\,\underline{u}'\underline{R}(t)\,\underline{u}\,, \qquad x_{n+1}(t_o) = 0\,, \tag{13}$$

and

$$x_{n+1}(T) = G \le E$$

one can proceed to form the Hamiltonian:

$$H = \lambda_o \frac{1}{2}\,\underline{x}'(t)\,\underline{Q}(t)\,\underline{x}(t) + \underline{\lambda}'\underline{A}(t)\,\underline{x}(t) + \underline{\lambda}'\underline{B}(t)\underline{u} + \lambda_{n+1}\frac{1}{2}\,\underline{u}'\underline{R}(t)\,\underline{u} \tag{14}$$

where

$$\underline{\lambda} \text{ is } n\times 1\,, \qquad \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix},$$

and $\dot{x}_o = \frac{1}{2}\,\underline{x}'(t)\,\underline{Q}(t)\,\underline{x}(t)$

Minimizing the Hamiltonian, one obtains

$$\underline{u} = -\frac{\underline{R}^{-1}(t)\,\underline{B}'(t)\,\underline{\lambda}(t)}{\lambda_{n+1}} \tag{15}$$

From the equations

$$\dot{\lambda}_i = -\frac{\partial H}{\partial x_i}\,, \qquad i = 0,\ \ldots,\ n+1$$

one obtains

$$\dot{\lambda}_o = 0\,, \qquad \dot{\underline{\lambda}} = -\underline{Q}(t)\,\underline{x} - \underline{A}'(t)\,\underline{\lambda}\,, \tag{16}$$

and

$$\dot{\lambda}_{n+1} = 0\,.$$

One then has

$$\begin{bmatrix} \dot{\underline{x}} \\ \dot{\underline{\lambda}} \end{bmatrix} = \begin{bmatrix} \underline{A}(t) & -\frac{1}{\lambda_{n+1}}\,\underline{B}(t)\,\underline{R}^{-1}(t)\,\underline{B}'(t) \\ -\underline{Q}(t) & -\underline{A}'(t) \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{\lambda} \end{bmatrix} \tag{17}$$

From transversality conditions, $\underline{\lambda}(T) = \underline{F}\,\underline{X}(T)$

From here one then can obtain (cf Kalman [1], Athans and Falb [3])

$$\underline{\lambda}(t) = \underline{K}(t)\ \underline{x}(t) \tag{18}$$

which makes the control law

$$\underline{u} = -\frac{1}{\lambda_{n+1}}\ \underline{R}^{-1}(t)\ \underline{B}'(t)\ \underline{K}(t)\ \underline{x}(t) \tag{19}$$

where the n×n matrix $\underline{K}(t)$ satisfies the differential equation

$$\dot{\underline{K}}(t) = -\underline{K}(t)\ \underline{A}(t) - \underline{A}'(t)\ \underline{K}(t) + \frac{\underline{K}(t)\underline{B}(t)\underline{R}^{-1}(t)\underline{B}(t)\underline{K}(t)}{\lambda_{n+1}} - \underline{Q}(t) \tag{20}$$

with boundary condition

$$\underline{K}(T) = \underline{F}$$

The unknown, λ_{n+1}, must be chosen so that the energy inequality constraint is satisfied. Some iterative procedure is required. Before deciding on an iterative procedure, certain observations should be made from equations 19 and 20. It is clear from these equations and equations 4 and 7 that one would have gotten this same result by taking as a cost function

$$J = \frac{1}{2}\ \underline{x}'(T)\ \underline{F}\ \underline{x}(T) + \frac{1}{2}\int_{t_o}^{T}[x'(t)\underline{Q}(t)\underline{x}(t) + \underline{u}'(t)\hat{\underline{R}}(t)\underline{u}(t)]dt, \tag{21}$$

where

$$\hat{\underline{R}}(t) = \lambda_{n+1}\ \underline{R}(t)\ ,$$

and imposing no constraints. We see that for a given $\underline{R}(t)$, increasing λ_{n+1} increases $\hat{R}(t)$ which causes the control energy of equation 21 to be weighted heavier and thus less control to be used. This causes

$$G = \frac{1}{2}\int_{t_o}^{T} \underline{u}'(t)\ \underline{R}(t)\ \underline{u}(t)\ dt \tag{22}$$

to be less as λ_{n+1} is made greater.

We thus argue heuristically that the curve of G vs λ_{n+1} is monotonically decreasing. See figure 1. This information is helpful in iterating to find the λ_{n+1} which causes G to be equal to E. It also seems plausible that the equality of the constraint should be used, for taking a $G < E$ corresponds to $\lambda_{n+1} > \lambda_{n+1_E}$ which means that the control energy of equation 21 is being weighted heavier than is necessary to satisfy the constraint. This corresponds to weighting error energy less than is necessary and causes error energy to be greater than would be the case if all available control energy were used.

III. NUMERICAL RESULTS

There are several reasons for wanting numerical results. First, we want to investigate the amount of computation and the number of iterations required for solutions. We also wish to know if limiting control energy, $\frac{1}{2}\int_{t_o}^{T}\underline{u}'(t)\underline{R}(t)\underline{u}(t)dt$, tends to limit $|u|_{max}$. Since the λ_{n+1} required to satisfy the energy constraint will vary with initial conditions, we wish to know if there is a functional relationship between λ_{n+1} and $\underline{x}(t_o)$. This is important since λ_{n+1} must be known to solve the problem. Iterating to determine λ_{n+1} may take too long for real time computation in which case, one would either have to store λ_{n+1} as a function of $\underline{x}(t_o)$ or have some way of determining λ_{n+1} given $\underline{x}(t_o)$. Once λ_{n+1} is known, one can integrate the differential equation of $\underline{K}(t)$ backwards and either store $\underline{K}(t)$ along the way or take $\underline{K}(t_o)$ and integrate the $\underline{K}(t)$ differential forward as the process takes place.

Two systems have been investigated: the harmonic oscillator,

$$\ddot{x} + x = u , \tag{23}$$

and a single integrator with damping,

$$\ddot{x} + \dot{x} = u . \tag{24}$$

In each case the control interval, $T - t_o$, was taken to be ten seconds, the control energy constraint value, E, was taken to be 1.5, $\underline{F} = .2\,\underline{I}$, and $\underline{Q} = .02\,\underline{I}$ ($\underline{I}$ is the identity matrix).

The procedure used was to first take λ_{n+1} to be unity, integrate the $\underline{K}(t)$ equation backwards storing the values of $\underline{K}(t)$, then integrate the equations of motion forward using the control law of equation (19).

A simple hyperbola was used as an approximation to the curve of figure 1. This gave the next value of λ_{n+1} to be:

$$\lambda_{n+1_{k+1}} = \frac{G_K}{E} \lambda_{n+1_k} \tag{25}$$

This procedure was used until an overshoot in G occurred, i.e., G crossed the value E. Then a linear interpolation based on the last two sets of data was used.

The average number of iterations required to bring G within .1 percent of E was 5.41 for the harmonic oscillator and 6.68 for the single integrator with damping. The computation times per iteration using a Burroughs B5500 computer averaged 0.54 and 0.53 minutes respectively for the two systems.

To show the effect of the energy constraint on $|u|_{max}$, figures 2 and 3 contain plots of $|u|_{max}$ versus initial conditions with $\lambda_{n+1} = 1$ and also with λ_{n+1} taking on the value required to satisfy the energy constraints.

The average value of $|u|_{max}$ and its variance for $\lambda_{n+1} = 1$ and for λ_{n+1} taking on the value required to satisfy the energy constraint are listed with these figures. These data strongly illustrate that constraining the control energy tends to constrain $|u|_{max}$. If a smaller value of $|u|_{max}$ was desired, then one would have to reduce E.

To study the behavior of λ_{n+1} as a function of initial conditions, the initial conditions were plotted in the $x_1 - x_2$ plane and the value of λ_{n+1} required to satisfy the energy constraint for each initial condition written in at that point in the $x_1 - x_2$ plane. The lines of constant λ_{n+1} were constructed. Figures 4 and 5 show these plots for each system and also give analytic expressions which one could use to compute approximate values of λ_{n+1} given $\underline{x}(t_o)$. These expressions are empirical and do not give the exact value of λ_{n+1} desired; however, the deviation between the exact and the approximate values of λ_{n+1} as listed in Tables I and II, which summarize the numerical results, are seen to be very small demonstrating the closeness of the empirical fit. The important point is that for the systems studied, there is an orderly behavior of λ_{n+1} as a function of initial state, $\underline{x}(t_o)$, and that one can obtain an approximate value for λ_{n+1} analytically given $\underline{x}(t_o)$.

It is interesting to note from Tables I and II that the empirical expression for $\lambda_{n+1}(\underline{x}(t_o))$ fits most poorly for initial states which are close to the origin. Also, $|u|_{max}$ takes on its largest values for these states. This is fortunate, for if the system state is already close to its desired value, then it is not as critical that one use up all available energy to get it closer. Thus one could choose a λ_{n+1} considerably larger than the one calculated from the empirical expression and in this way ensure that the energy constraint is not violated and that $|u|_{max}$ does not get out of hand if this is a consideration. Furthermore the inaccuracies in the calculated λ_{n+1} are such that if this value of λ_{n+1} is taken, less than E amount of energy will be used.

Some sample trajectories are shown in Figures 6 and 7. One can compare the behavior of the systems where the exact value of λ_{n+1} was used and where the approximate value was used.

IV. CONCLUSIONS

The quadratic control regulator problem with an energy constraint has been studied. The theoretical development yielded the Riccati equation with an unknown parameter, λ_{n+1}. Because of this parameter, implementation of the result is not as straightforward as is Kalman's result where there are no constraints. However, the problem with constraints is more realistic. Numerical solutions were obtained for two problems via a straightforward iterative technique. It was found for these problems that limiting the control energy tended to limit $|u|_{max}$ providing an indirect method for taking into account a control amplitude constraint. The parameter λ_{n+1}, was seen to have a regular behavior as a function of the initial condition of the state which means that one could calculate an approximate value for λ_{n+1} given $\underline{x}(t_o)$ instead of having to store values of λ_{n+1} for all possible initial conditions. Implementation would then require: given $\underline{x}(t_o)$, calculate $_{n+1}$, integrate the Riccati equation backwards, either store $\underline{K}(t)$ during the backward integration or having found $\underline{K}(t_o)$, integrate the Riccati equation forward in real time as the process to be controlled takes place.

V. ACKNOWLEDGEMENT

The authors wish to thank Professor Michael Athans of the Electrical Engineering Department, Massachusetts Institute of Technology, for suggesting this problem as an area of possible research.

REFERENCES

1. Kalman, R.E., "Contributations to the Theory of Optimal Control", Boletin de la Sociedad Matematica Mexicana, 1960.
2. Newton, G.C., Jr., "Compensation of Feedback Control Systems Subject to Saturation", Jnl of Franklin Institute, 254, 281-286, 391-413, 1952.
3. Athans and Falb, Optimal Control An Introduction to the Theory and Its Applications, McGraw-Hill Book Co., Inc., N.Y., 1966, p. 757-761.

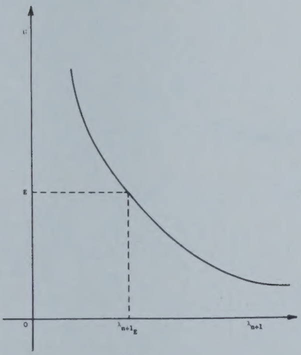


FIGURE 1

Expected behavior of G as a function of λ_{n+1}

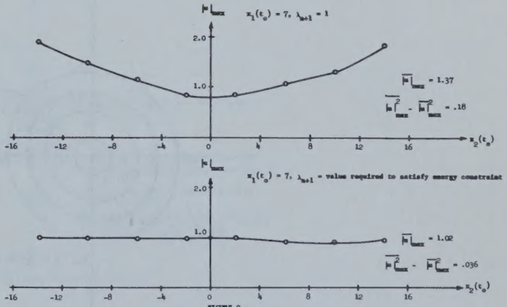


FIGURE 2

$|u|_{max}$ vs initial conditions - Harmonic Oscillator

Note: The average values and variances are for all the initial conditions listed in Table 1, not just the ones shown in this figure.

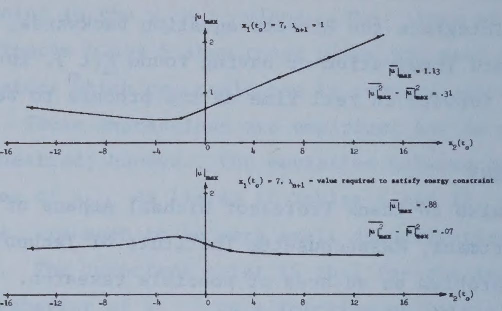


FIGURE 3

$|u|_{max}$ vs initial conditions - Single Integrator with Damping

Note: The average values and variances are for all the initial conditions listed in Table III, not just the ones shown in this figure.

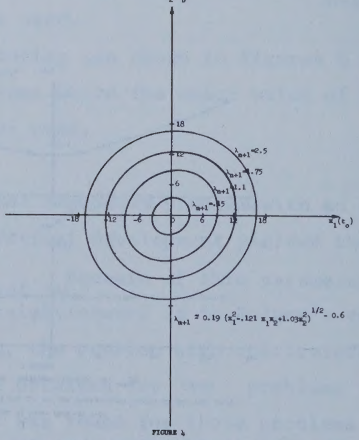
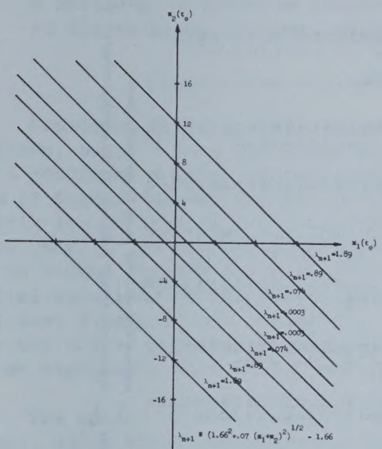


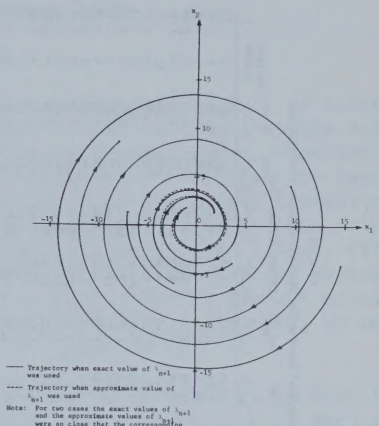
FIGURE 4

Lines of constant λ_{n+1} as a function of initial conditions - Harmonic Oscillator



$\lambda_{n+1} = (1.66^2 + .07\,(x_1+x_2)^2)^{1/2} - 1.66$

FIGURE 5
Lines of constant λ_{n+1} as a function of initial conditions - single integrator with damping

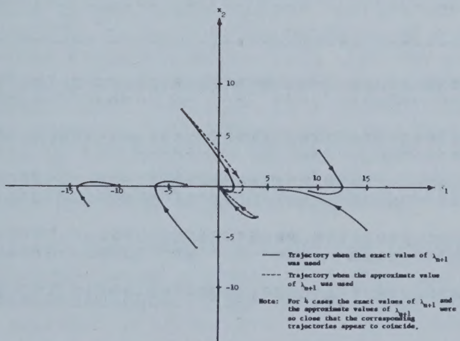


—— Trajectory when exact value of λ_{n+1} was used

---- Trajectory when approximate value of λ_{n+1} was used

Note: For two cases the exact values of λ_{n+1} and the approximate values of λ_{n+1} were so close that the corresponding trajectories appear to coincide.

FIGURE 6
Trajectories for Harmonic Oscillator



—— Trajectory when the exact value of λ_{n+1} was used

---- Trajectory when the approximate value of λ_{n+1} was used

Note: For 4 cases the exact values of λ_{n+1} and the approximate values of λ_{n+1} were so close that the corresponding trajectories appear to coincide.

FIGURE 7
Trajectories for single integrator with damping

Table I

Numerical Results for Harmonic Oscillator

$x_1(0)$	$x_2(0)$	2XG	$\|u\|_{max}$	2XCost	2XCost	$\|u\|_{max}$	λ_f	No.Iter-ations	$\hat{\lambda}$
		$\lambda_{n+1}=1$	$\lambda_{n+1}=1$	$\lambda_{n+1}=1$	$\lambda_{n+1}=\lambda_f$	$\lambda_{n+1}=\lambda_f$			
0	4	.67	.52	1.40	.43	1.45	.15	8	.17
0	-4	.67	.52	1.40	.43	1.45	.15	8	.17
4	12	6.58	1.57	14.51	19.23	1.02	1.80	5	1.79
0	-12	6.05	1.55	12.56	16.51	1.06	1.71	5	1.79
2	14	8.32	1.82	17.76	25.3	1.05	2.11	5	2.10
2	10	4.31	1.30	9.30	10.80	1.06	1.33	6	1.34
2	-6	1.71	.79	3.37	2.39	1.09	.59	6	.64
2	-10	4.42	1.29	8.87	10.5	1.04	1.36	6	1.38
4	15	9.95	1.95	21.70	32.0	1.03	2.36	5	2.34
2	-14	8.49	1.80	17.16	25.0	1.03	2.15	5	2.14
15	4	9.84	1.79	22.22	32.29	.97	2.32	5	2.30
15	-4	10.19	1.81	20.95	31.70	.96	2.41	5	2.39
15	0	9.35	1.74	20.19	29.44	.97	2.26	5	2.25
0	-8	2.69	1.03	5.58	5.28	1.10	.91	4	.94
4	0	.67	.46	1.44	.47	1.14	.15	8	.16
0	8	2.69	1.03	5.58	5.29	1.10	.91	4	.94
7	14	9.99	1.84	22.5	32.88	.96	2.36	5	2.33
7	10	6.03	1.34	13.86	17.76	.93	1.68	5	1.67
7	6	3.43	1.04	7.98	8.43	.97	1.11	5	1.10
7	2	2.16	.84	4.89	4.17	1.00	.75	5	.76
7	-2	2.24	.85	4.60	3.93	.99	.78	5	.81
7	-6	3.67	1.09	7.09	7.82	.98	1.18	6	1.21
4	-12	6.86	1.58	13.49	18.68	1.01	1.87	5	1.87
4	4	1.29	.63	3.00	1.87	1.00	.44	8	.45
4	-4	1.38	.67	2.66	1.57	1.05	.46	7	.51
4	-15	10.3	1.95	20.43	31.42	1.01	2.44	5	2.43
0	-15	9.46	1.94	19.63	29.12	1.04	2.30	5	2.29
2	6	1.64	.78	3.63	2.60	1.12	.57	6	.60
4	8	3.26	1.05	7.36	7.63	1.00	1.07	5	1.07
4	-8	3.45	1.10	6.68	7.15	1.02	1.12	5	1.15
7	-10	6.44	1.47	12.38	16.93	.98	1.79	5	1.80
7	-14	10.56	1.92	20.46	31.93	.98	2.49	5	2,47
10	12	9.85	1.73	22.81	32.90	.94	2.33	5	2.30
10	8	6.61	1.44	15.41	20.2	.96	1.79	5	1.77
10	4	4.71	1.24	10.79	12.81	.98	1.41	6	1.40
8	8	5.16	1.26	12.0	14.64	.95	1.51	6	1.50
10	0	4.15	1.16	8.97	10.28	.98	1.29	6	1.30
10	-4	4.94	1.26	9.94	10.29	.97	1.48	6	1.49
10	-8	7.08	1.51	13.71	19.24	.95	1.91	4	1.91
10	-12	10.56	1.87	20.26	31.73	.97	2.49	5	2.48
13	10	10.84	1.85	25.27	37.07	.95	2.47	5	2.44
13	6	8.30	1.64	19.13	26.59	.97	2.08	4	2.06
13	2	7.11	1.52	15.79	21.34	.97	1.89	5	1.87
13	-2	7.26	1.53	15.24	21.04	.97	1.92	4	1.92
13	-6	8.76	1.68	17.48	25.78	.96	2.19	5	2.18
13	-10	11.61	1.94	22.51	35.85	.96	2.63	6	2.62
5	5	2.02	.79	4.68	3.86	.98	.71	5	.71
0	15	9.46	1.94	19.63	29.12	1.04	2.30	5	2.29
0	12	6.05	1.55	12.56	16.51	1.06	1.71	5	1.71

λ_f is the value of λ_{n+1} required to satisfy the energy equality constraint.

$\hat{\lambda}$ is the value of λ_{n+1} obtained from initial conditions via the empirical formula.

Table II

Numerical Results for Single Integrator with Damping

$x_1(0)$	$x_2(0)$	2XG	$\|u\|_{max}$	2XCost	2XCost	$\|u\|_{max}$	λ_f	No.Iter-ations	$\hat{\lambda}$
		$\lambda_{n+1}=1$	$\lambda_{n+1}=1$	$\lambda_{n+1}=1$	$\lambda_{n+1}=\lambda_f$	$\lambda_{n+1}=\lambda_f$			
7	-14	2.05	.71	3.38	2.62	.9	.64	5	.83
4	-15	5.07	1.11	7.29	9.94	.82	1.63	4	1.69
0	-12	6.04	1.21	8.49	12.64	.81	1.89	5	1.92
2	14	10.74	1.61	15.13	2.79	.79	2.89	7	2.88
2	10	6.04	1.21	8.53	12.68	.81	1.89	5	1.92
2	-6	.67	.40	.98	.25	1.48	.07	13	.31
2	-10	2.68	.81	3.81	3.52	.86	.89	3	1.02
4	15	15.14	1.92	21.4	43.5	.78	3.65	8	3.63
2	-14	6.04	1.21	8.53	12.68	.81	1.89	5	1.92
4	12	10.74	1.61	15.25	28.04	.79	2.89	7	2.88
15	-4	5.08	1.11	9.38	12.03	.82	1.63	4	1.69
15	0	9.44	1.51	15.52	25.8	.79	2.64	6	2.64
13	10	22.2	2.32	32.9	70.9	.78	4.65	10	4.64
13	6	15.1	1.92	22.97	44.99	.78	3.65	8	3.63
13	2	9.44	1.51	14.96	25.2	.79	2.64	6	2.64
13	-2	5.07	1.11	8.82	11.47	.82	1.63	4	1.69
13	-6	2.06	.71	4.58	3.82	.90	.64	5	.83
15	4	15.1	1.92	23.5	45.6	.78	2.65	8	3.63
0	-8	2.68	.81	3.77	3.48	.86	.89	3	1.02
10	12	20.3	2.22	29.5	63.1	.78	4.40	9	4.39
10	8	13.6	1.82	20.1	38.8	.79	3.39	7	3.38
10	4	8.22	1.41	12.6	20.5	.81	2.39	6	2.39
8	8	10.74	1.61	15.7	28.5	.79	2.89	7	2.88
10	0	4.19	1.01	6.9	8.3	.83	1.38	4	1.46
10	-4	1.51	.61	3.12	2.16	.98	.41	8	.64
0	-15	9.44	1.51	13.3	23.5	.79	2.64	6	2.64
2	6	2.68	.81	3.81	3.52	.86	.89	3	1.02
4	-8	.67	.40	1.10	.37	1.48	.07	13	.31
7	10	12.1	1.72	17.5	33.1	.79	3.14	7	3.13
7	6	7.09	1.31	10.5	16.4	.80	2.14	6	2.15
7	2	3.40	.91	5.27	5.69	.84	1.14	3	1.24
4	8	6.04	1.21	8.65	12.8	.81	1.89	5	1.92
5	5	4.19	1.01	6.15	7.55	.83	1.38	4	1.46
0	15	9.44	1.51	13.3	23.5	.79	2.64	6	2.64
0	12	6.04	1.21	8.49	12.64	.81	1.89	5	1.92
0	8	2.68	.81	3.77	3.48	.86	.89	3	1.02
4	-12	2.68	.81	3.93	3.64	.86	.89	3	1.02
4	4	2.68	.81	3.93	3.64	.86	.89	3	1.02
4	0	.67	.40	1.1	.37	1.48	.074	13	.31
7	-2	1.05	.50	1.96	1.03	1.14	.20	11	.46
0	-4	.67	.40	.94	.21	1.48	.074	13	.31
0	4	.67	.40	.94	.21	1.48	.074	13	.31

λ_f is the value of λ_{n+1} required to satisfy the energy equality constraint.

$\hat{\lambda}$ is the value of λ_{n+1} obtained from initial conditions via the empirical formula.

ASYMPTOTIC SOLUTION OF THE MATRIX RICCATI EQUATION OF OPTIMAL CONTROL

JOHN J. O'DONNELL
Applied Research Laboratory
Sylvania Electronic Systems
A Division of Sylvania Electric Products Inc.
40 Sylvan Road, Waltham, Massachusetts, 02154

ABSTRACT

Construction of a feedback matrix for the optimal control of linear systems, subject to a quadratic performance criterion, requires the solution of a nonlinear matrix differential equation of Riccati type. When the system is time invariant and the optimization period is infinite, the feedback matrix is obtained from the asymptotic solution of the matrix Riccati equation (MRE). The conventional approach to obtaining the asymptotic solution is to reduce the MRE to an auxiliary system of homogeneous, linear, differential equations of the first order and integrate this auxiliary system until some steady-state is reached. In this paper there is derived an algorithm for the direct computation of the asymptotic solution of the MRE, by means of an eigenvalue-eigenvector analysis of a single matrix.

The matrix of the auxiliary linear system for the MRE is of Hamiltonian type. Using the eigenvectors corresponding to the positive real part eigenvalues of this Hamiltonian matrix, two matrices are formed whose product is shown to be the asymptotic solution of the MRE. In addition, it is demonstrated that the eigenvalues and eigenvectors of the optimal closed-loop system matrix are contained in those of the Hamiltonian matrix; hence the closed-loop properties (transient and frequency responses) of the optimal system are easily calculated. The numerical merits of this new method and the conventional method are compared.

INTRODUCTION

The solution of linear, least-squares control and estimation problems is known to be reducible to finding the solution of a first-order, nonlinear, matrix equation — the Riccati equation [1]. For the general, time-varying situation there is no recourse but numerical integration to obtain a solution; in the constant-coefficient case there is a possibility for analytical solution for low order problems. When the system is time-invariant, and the optimization period is infinite, the Riccati equation has an asymptotic or stationary solution: a constant matrix. Analytically, this constant matrix is the solution of a nonlinear algebraic equation.

From a numerical standpoint, the nonlinear, algebraic equation system constitutes an ill-conditioned problem. The practical approach is simply to find the asymptotic solution of the Riccati equation. One standard method for finding this solution is to convert the n-dimensional Riccati equation to a 2n-dimensional homogeneous, linear differential equation whose solution is found by calculating the matrix exponential [2]. In this paper an alternative approach is derived: the asymptotic solution may be found by an eigenvalue-eigenvector analysis of the Riccati system matrices.

MATRIX RICCATI EQUATION

Both the control and the estimation problem reduce to finding the solution of the same matrix Riccati equation. Only the control results will be discussed in this note; the estimation results follow by duality. The optimal, least-squares control problem and its solution are summarized below[*] (see [1]).

Plant:
$$\dot{\underline{x}} = \underline{F}\,\underline{x} + \underline{D}\,\underline{v} \tag{1}$$
$$\underline{y} = \underline{E}\,\underline{x} \tag{2}$$

Performance Index:
$$J = \int_0^\infty \left[\underline{y}^T \underline{M}\,\underline{y} + \underline{v}^T \underline{N}\,\underline{v}\right] dt \quad ; \quad \underline{M} \text{ and } \underline{N} \text{ positive definite} \tag{3}$$

Optimal Control:
$$\underline{v}^* = -\underline{N}^{-1}\underline{D}^T\underline{L}(\infty)\underline{x} \tag{4}$$

Loss Equation: (Riccati Equation)
$$\dot{\underline{L}} = \underline{L}\,\underline{F} + \underline{F}^T\underline{L} - \underline{L}\,\underline{D}\,\underline{N}^{-1}\underline{D}^T\underline{L} + \underline{E}^T\underline{M}\,\underline{E} \tag{5}$$
$$\underline{L}(0) = \underline{E}^T\underline{M}\,\underline{E} \tag{6}$$
$$\underline{L}(\infty) = \lim_{t \to \infty} \underline{L}(t) \tag{7}$$

The matrix $\underline{L}$, which is the solution of the loss, or Riccati, equation is a positive-definite matrix which is the only unknown quantity in the expression for the feedback matrix, $\underline{N}^{-1}\underline{D}^T\underline{L}(\infty)$. To find it, it is necessary to convert the loss equation to a linear system by the substitution of variable
$$\underline{L}(t) = \underline{Z}(t)\underline{Y}^{-1}(t) \quad . \tag{8}$$

The initial condition (6) becomes
$$\underline{Z}(0) = \underline{E}^T\underline{M}\,\underline{E} \quad , \qquad \underline{Y}(0) = \underline{I} \quad . \tag{9}$$

Substitution of (8) into (5) yields two linear differential equations which can be combined into the single, 2n-dimensional equation
$$\left[\begin{array}{c}\dot{\underline{Y}} \\ \hline \dot{\underline{Z}}\end{array}\right] = \left[\begin{array}{c|c} -\underline{F} & \underline{D}\,\underline{N}^{-1}\underline{D}^T \\ \hline \underline{E}^T\underline{M}\,\underline{E} & \underline{F}^T \end{array}\right]\left[\begin{array}{c}\underline{Y} \\ \hline \underline{Z}\end{array}\right] = \underline{H}\left[\begin{array}{c}\underline{Y} \\ \hline \underline{Z}\end{array}\right] \quad . \tag{10}$$

Then Equation (1) has the solution, using (9),
$$\left[\begin{array}{c}\underline{Y} \\ \hline \underline{Z}\end{array}\right] = e^{\underline{H}t}\left[\begin{array}{c}\underline{I} \\ \hline \underline{E}^T\underline{M}\,\underline{E}\end{array}\right] \quad . \tag{11}$$

ASYMPTOTIC SOLUTION OF THE RICCATI EQUATION

Equations (12) and (8) completely specify the solution to the loss equation, (5), for all finite t. Additional constraints must be placed on the problem statement if the asymptotic solution $\underline{L}(\infty)$ is to exist.

---

[*] The problem stated is actually the optimal regulator problem. The servo problems, however, require solution of the same Riccati equation.

Kalman has proven [1] [2] that if the plant is completely controllable and completely observable then the Riccati equation possesses a unique solution. The purpose of this section will be to obtain an exact algebraic statement for the form of this unique $\underline{L}(\infty)$.

In the Appendix it is proved that

$$\underline{P}^{-1}\underline{H}\,\underline{P} = \left[\begin{array}{c|c} \underline{\Delta} & \underline{0} \\ \hline \underline{0} & -\underline{\Delta} \end{array}\right] = \underline{\Delta}_1 \tag{12}$$

where

$$\underline{\Delta} = \mathrm{diag}\left[\lambda_1, \lambda_2, \cdots, \lambda_n\right] , \qquad \mathrm{Re}\left[\lambda_i\right] > 0 , \tag{13}$$

$$\underline{P}^{-1} = -\underline{J}\,\underline{P}^T\underline{J} , \qquad \underline{P} = \left[\underline{\boldsymbol{\xi}}_1, \cdots, \underline{\boldsymbol{\xi}}_n, \underline{\zeta}_1, \cdots, \underline{\zeta}_n\right] , \tag{14}$$

$$\underline{J} = \left[\begin{array}{c|c} \underline{0} & -\underline{I} \\ \hline \underline{I} & \underline{0} \end{array}\right] , \qquad \begin{aligned} \underline{H}\,\underline{\boldsymbol{\xi}}_i &= \lambda_i\underline{\boldsymbol{\xi}}_i , \\ \underline{H}\,\underline{\zeta}_i &= -\lambda_i\underline{\zeta}_i . \end{aligned} \tag{15}$$

The assumption has been made that the eigenvalues of $\underline{H}$ are distinct with nonzero real parts. The nonzero character is necessary for ensuring that the closed-loop optimal system is stable (see Eq. (30)). For numerical, as well as engineering reasons, the assumption of distinct eigenvalues is no restriction, and in fact is realistic. Note that it may be true that λ_j and λ_{j+1}, say, are complex conjugates with positive real parts. In that case, the corresponding eigenvectors have complex elements. In general a computer program to determine eigenvalues and eigenvectors should have provisions for computing complex numbers.

It is of course true that each eigenvector is unique in direction only and can have any magnitude, i.e.,

$$\underline{H}(\alpha_i\underline{\boldsymbol{\xi}}_i) = \lambda_i(\alpha_i\underline{\boldsymbol{\xi}}_i) \qquad \text{and} \qquad \underline{H}\,\underline{\boldsymbol{\xi}}_i = \lambda_i\underline{\boldsymbol{\xi}}_i$$

are the same. Consequently, arbitrary restrictions on the eigenvector magnitudes are necessary to insure complete uniqueness. Typical restrictions are to require all eigenvectors to have unit length, or to require that some element of the eigenvector be unity. For the derivation which follows in this paper, no restriction on the magnitude of the eigenvectors is essential. Hence it may be thought that the final result, the asymptotic solution to the Riccati equation expressed in terms of the eigenvectors of $\underline{H}$, could not be unique. It will be demonstrated, after Eq. (22), that the result is unique regardless of the nonuniqueness of the magnitude of the eigenvectors.

From Eq. (15), if

$$\underline{P} = \left[\begin{array}{c|c} \underline{P}_{11} & \underline{P}_{12} \\ \hline \underline{P}_{21} & \underline{P}_{22} \end{array}\right] \quad \text{(a)} , \quad \text{then} \quad \underline{P}^{-1} = \left[\begin{array}{c|c} \underline{P}_{22}^T & -\underline{P}_{12}^T \\ \hline -\underline{P}_{21}^T & \underline{P}_{11}^T \end{array}\right] \quad \text{(b)} \tag{16}$$

From Eq. (12) $\underline{H} = \underline{P}\,\underline{\Lambda}_1\underline{P}^{-1}$; by definition of the matrix exponential there results

$$e^{\underline{H}t} = \underline{P}\, e^{\underline{\Lambda}_1 t}\, \underline{P}^{-1} \quad . \tag{17}$$

Using the partitioned forms for $\underline{P}$ (Eq. (16a)), $\underline{P}^{-1}$ (Eq. (16b)), and $\underline{\Lambda}_1$ (Eq. (12)), Eq. (17) becomes

$$e^{\underline{H}t} = \left[\begin{array}{c|c} \underline{P}_{11} & \underline{P}_{12} \\ \hline \underline{P}_{21} & \underline{P}_{22} \end{array}\right] \left[\begin{array}{c|c} e^{\underline{\Lambda}t} & \underline{0} \\ \hline \underline{0} & e^{-\underline{\Lambda}t} \end{array}\right] \left[\begin{array}{c|c} \underline{P}_{22}^T & -\underline{P}_{12}^T \\ \hline -\underline{P}_{21}^T & \underline{P}_{11}^T \end{array}\right] ,$$

or

$$e^{\underline{H}t} = \left[\begin{array}{c|c} \underline{P}_{11}e^{\underline{\Lambda}t}\underline{P}_{22}^T - \underline{P}_{12}e^{-\underline{\Lambda}t}\underline{P}_{21}^T & \underline{P}_{12}e^{-\underline{\Lambda}t}\underline{P}_{11}^T - \underline{P}_{11}e^{\underline{\Lambda}t}\underline{P}_{12}^T \\ \hline \underline{P}_{21}e^{\underline{\Lambda}t}\underline{P}_{22}^T - \underline{P}_{22}e^{-\underline{\Lambda}t}\underline{P}_{21}^T & \underline{P}_{22}e^{-\underline{\Lambda}t}\underline{P}_{11}^T - \underline{P}_{21}e^{\underline{\Lambda}t}\underline{P}_{12}^T \end{array}\right] . \tag{18}$$

Using Eqs. (8), (11), and (18) the expression for $\underline{L}(t)$ becomes

$$\begin{aligned} \underline{L}(t) &= \underline{Z}(t)\underline{Y}^{-1}(t) \\ &= \left[\underline{P}_{21}e^{\underline{\Lambda}t}\underline{P}_{22}^T - \underline{P}_{22}e^{-\underline{\Lambda}t}\underline{P}_{21}^T + (\underline{P}_{22}e^{-\underline{\Lambda}t}\underline{P}_{11}^T - \underline{P}_{21}e^{\underline{\Lambda}t}\underline{P}_{12}^T)\underline{E}^T\underline{M}\,\underline{E}\right] \\ &\quad \cdot \left[\underline{P}_{11}e^{\underline{\Lambda}t}\underline{P}_{22}^T - \underline{P}_{12}e^{-\underline{\Lambda}t}\underline{P}_{21}^T + (\underline{P}_{12}e^{-\underline{\Lambda}t}\underline{P}_{11}^T - \underline{P}_{11}e^{\underline{\Lambda}t}\underline{P}_{12}^T)\underline{E}^T\underline{M}\,\underline{E}\right]^{-1} . \end{aligned} \tag{19}$$

The asymptotic solution of the Riccati equation is now to be evaluated from Eq. (19) by forming the limit as t approaches infinity. A mathematically rigorous derivation of this limit is, however, difficult because neither $\underline{Z}(t)$ nor $\underline{Y}(t)$ possesses a limit, as may be seen from Eq. (19). What Kalman [1] [2] has proved, under the assumptions of complete controllability and observability, is that the product in Eq. (19) possesses a limit as $t \to \infty$. Consequently the following argument is nonrigorous and yields a result (which is stated in the theorem at the end of this section) which must be qualified by certain undoubtedly superfluous assumptions.

Under the assumptions of complete controllability and observability, $\underline{L}(t)$ exists for all t [1] [2]. Hence, $\underline{Y}(t)$ must be nonsingular for all t. As t becomes large the terms in Eq. (19) which contain $e^{\underline{\Lambda}t}$ as a factor will certainly dominate* those containing $e^{-\underline{\Lambda}t}$ as a factor. Hence, for t sufficiently large, it is justifiable from the numerical viewpoint to represent $\underline{L}(t)$ as

$$\underline{L}(t) \underset{t\to\infty}{=} \left[\underline{P}_{21}e^{\underline{\Lambda}t}(\underline{P}_{22}^T - \underline{P}_{12}^T\underline{E}^T\underline{M}\,\underline{E})\right]\left[\underline{P}_{11}e^{\underline{\Lambda}t}(\underline{P}_{22}^T - \underline{P}_{12}^T\underline{E}^T\underline{M}\,\underline{E})\right]^{-1} , \tag{20}$$

---

*Recall that all diagonal elements of $\underline{\Lambda}$ have positive real parts. If one diagonal element of $\underline{\Lambda}$ were to have a zero real part, the argument leading to Eq. (22) could not be justified.

or

$$\underline{L}(t) \underset{t\to\infty}{=} \underline{P}_{21}e^{\underline{\Lambda}t}(\underline{P}_{22}^T - \underline{P}_{12}^T\underline{E}^T\underline{M}\ \underline{E})(\underline{P}_{22}^T - \underline{P}_{12}^T\underline{E}^T\underline{M}\ \underline{E})^{-1}e^{-\underline{\Lambda}t}\underline{P}_{11}^{-1} \quad . \tag{21}$$

Where the indicated inverses in Eq. (21) are assumed to exist. Performance of the obvious cancellations in Eq. (21) leads to the asymptotic value for $\underline{L}$,

$$\underline{L}(\infty) = \underline{P}_{21}\underline{P}_{11}^{-1} \quad . \tag{22}$$

It is now easy to show that the nonuniqueness of the magnitudes of the eigenvectors of $\underline{H}$ does not imply the nonuniqueness of the result in Eq. (22). Another possible set of eigenvectors for $\underline{H}$ is, for α_i finite and nonzero,

$$\left[\alpha_1\underline{\xi}_1,\cdots,\alpha_n\underline{\xi}_n\right] = \begin{bmatrix}\underline{P}_{11}\\ \hline \underline{P}_{21}\end{bmatrix}\underline{A} = \begin{bmatrix}\underline{P}_{11}\underline{A}\\ \hline \underline{P}_{21}\underline{A}\end{bmatrix} \quad ; \quad \underline{A} = \operatorname{diag}\left[\alpha_1,\cdots,\alpha_n\right] \quad .$$

With this new set of eigenvectors Eq. (22) becomes

$$\underline{L}(\infty) = (\underline{P}_{21}\underline{A})(\underline{P}_{11}\underline{A})^{-1} = \underline{P}_{21}\underline{P}_{11}^{-1} \quad .$$

Thus the result (22) is unaffected by the choice of magnitude normalization which is used for the eigenvectors.

It was necessary to assume, in the transition from Eq. (20) to Eq. (21), that the matrices $\underline{P}_{11}$ and $(\underline{P}_{22}^T - \underline{P}_{12}^T\underline{E}^T\underline{M}\ \underline{E})$ were nonsingular. The following computation will show that only the nonsingularity of $\underline{P}_{11}$ need be assumed when verifying that Eq. (22) does provide a stationary solution of the Riccati equation, (5). What will be demonstrated is that

$$\underline{P}_{21}\underline{P}_{11}^{-1}\underline{F} + \underline{F}^T\underline{P}_{21}\underline{P}_{11}^{-1} - \underline{P}_{21}\underline{P}_{11}^{-1}\underline{D}\ \underline{N}^{-1}\underline{D}^T\underline{P}_{21}\underline{P}_{11}^{-1} + \underline{E}^T\underline{M}\ \underline{E} \equiv \underline{0} \tag{23}$$

which is the result of combining Eqs. (5) and (22). In the following computation only the nonsingularity of $\underline{P}_{11}$ will be assumed. Substitution of $\underline{H}$, Eq. (10), and Eqs. (16a) and (16b) into Eq. (12) yields

$$\left[\begin{array}{c|c}\underline{P}_{22}^T & -\underline{P}_{12}^T\\ \hline -\underline{P}_{21}^T & \underline{P}_{11}^T\end{array}\right]\left[\begin{array}{c|c}-\underline{F}\ \underline{P}_{11} + \underline{D}\ \underline{N}^{-1}\underline{D}^T\underline{P}_{21} & -\underline{F}\ \underline{P}_{12} + \underline{D}\ \underline{N}^{-1}\underline{D}^T\underline{P}_{22}\\ \hline \underline{E}^T\underline{M}\ \underline{E}\ \underline{P}_{11} + \underline{F}^T\underline{P}_{21} & \underline{E}^T\underline{M}\ \underline{E}\ \underline{P}_{12} + \underline{F}^T\underline{P}_{22}\end{array}\right] = \left[\begin{array}{c|c}\underline{\Lambda} & \underline{0}\\ \hline \underline{0} & -\underline{\Lambda}\end{array}\right] . \tag{24}$$

Eq. (24) gives four matrix algebraic equations; select that one in the (2,1) position. Thus

$$-\underline{P}_{21}^T(-\underline{F}\ \underline{P}_{11} + \underline{D}\ \underline{N}^{-1}\underline{D}^T\underline{P}_{21}) + \underline{P}_{11}^T(\underline{E}^T\underline{M}\ \underline{E}\ \underline{P}_{11} + \underline{F}^T\underline{P}_{21}) = \underline{0} \quad . \tag{25}$$

When Eq. (25) is premultiplied by $(\underline{P}_{11}^{-1})^T$ and postmultiplied by $\underline{P}_{11}^{-1}$ there results

$$(\underline{P}_{11}^{-1})^T\underline{P}_{21}^T\underline{F} + \underline{F}^T\underline{P}_{21}\underline{P}_{11}^{-1} - (\underline{P}_{11}^{-1})^T\underline{P}_{21}^T\underline{D}\ \underline{N}^{-1}\underline{D}^T\underline{P}_{21}\underline{P}_{11}^{-1} + \underline{E}^T\underline{M}\ \underline{E} = \underline{0} \quad . \tag{26}$$

But $\underline{L}$ is symmetric, so that $\underline{L}(\infty) = \underline{P}_{21}\underline{P}_{11}^{-1} = (\underline{P}_{11}^{-1})^T\underline{P}_{21}^T$. Hence Eq. (26) becomes

$$\underline{P}_{21}\underline{P}_{11}^{-1}\underline{F} + \underline{F}^T\underline{P}_{21}\underline{P}_{11}^{-1} - \underline{P}_{21}\underline{P}_{11}^{-1}\underline{D}\,\underline{N}^{-1}\underline{D}^T\underline{P}_{21}\underline{P}_{11}^{-1} + \underline{E}^T\underline{M}\,\underline{E} = 0 \quad , \tag{27}$$

which was to be proved. The results of this section may be summarized in the following theorem.

THEOREM: Let $\underline{H}$ be the Hamiltonian matrix

$$\underline{H} = \left[\begin{array}{c|c} -\underline{F} & \underline{D}\,\underline{N}^{-1}\underline{D}^T \\ \hline \underline{E}^T\underline{M}\,\underline{E} & \underline{F}^T \end{array}\right]_{2n \times 2n}$$

with $\underline{M}$ and $\underline{N}$ positive definite matrices; and let $\underline{\xi}_i$ be the (right) eigenvector corresponding to the eigenvalue λ_i, with $\mathrm{Re}(\lambda_i) > 0$. If (i) all eigenvalues of $\underline{H}$ are distinct with nonzero real parts, and (ii) the matrix $\underline{P}_{11}$ is nonsingular, where

$$\left[\begin{array}{c} \underline{P}_{11} \\ \hline \underline{P}_{21} \end{array}\right] = \left[\underline{\xi}_1, \cdots, \underline{\xi}_n\right]_{2n \times n} \quad ,$$

then the Riccati equation

$$\dot{\underline{L}} = \underline{L}\,\underline{F} + \underline{F}^T\underline{L} - \underline{L}\,\underline{D}\,\underline{N}^{-1}\underline{D}^T\underline{L} + \underline{E}^T\underline{M}\,\underline{E}$$

has the asymptotic (stationary) solution

$$\underline{L}(\infty) = \underline{P}_{21}\underline{P}_{11}^{-1} .$$

As stated, the theorem has been proved by the development beginning at Eq. (19) and ending at Eq. (27). However, it is believed that the conditions of the theorem could be modified as stated in the following conjecture: CONJECTURE: If the matrices $[\underline{D}, \underline{F}\,\underline{D}, \cdots, \underline{F}^{n-1}\underline{D}]$ and $[\underline{E}^T, \underline{F}^T\underline{E}^T, \cdots, (\underline{F}^T)^{n-1}\underline{E}^T]$ are both of rank n then conditions (i) and (ii) of the above theorem will hold. The justification for advancing this conjecture is provided by the results of Kalman [1] [2]: i.e., complete controllability and complete observability of the plant imply both that the optimal closed-loop system is uniformly asymptotically stable, and that the asymptotic (stationary) solution of the Riccati equation must exist. The relationship between the eigenvalues of $\underline{H}$ and the poles of the optimal closed-loop system is explored in the next section.

The results of the above theorem were originally presented by this author in [5]. In [6] (which was actually published in August, 1966) Potter has independently come to the same conclusion although by an entirely different approach and with no special reference to the theory of least-squares optimal control.

POLES OF THE OPTIMAL CLOSED-LOOP SYSTEM

With $\underline{L}(\infty) = \underline{P}_{21}\underline{P}_{11}^{-1}$, the optimal control is (see Eq. (4))

$$\underline{v}^* = -\underline{N}^{-1}\underline{D}^T\underline{P}_{21}\underline{P}_{11}^{-1}\underline{x} \quad . \tag{28}$$

The optimal closed-loop regulating system is described by

$$\dot{\underline{x}} = (\underline{F} - \underline{D}\,\underline{N}^{-1}\underline{D}^T\underline{P}_{21}\underline{P}_{11}^{-1})\underline{x} \quad . \tag{29}$$

Information concerning the poles of the optimal closed-loop system, or the eigenvalues of the matrix in Eq. (33), may be obtained from studying Eq. (13) in the modified form $\underline{H}\,\underline{P} = \underline{P}\,\underline{\Lambda}_1$. The product matrix $\underline{H}\,\underline{P}$ is already displayed in Eq. (24). From inspection of Eq. (24) it is seen that the (1,1) element of $\underline{H}\,\underline{P}$ bears most resemblance to the matrix of Eq. (29). Hence, from Eq. (24) and $\underline{H}\,\underline{P} = \underline{P}\,\underline{\Lambda}_1$, there results $-\underline{F}\,\underline{P}_{11} + \underline{D}\,\underline{N}^{-1}\underline{D}^T\underline{P}_{21} = \underline{P}_{11}\underline{\Lambda}$. Factoring $\underline{P}_{11}$ gives

$$(\underline{F} - \underline{D}\,\underline{N}^{-1}\underline{D}^T\underline{P}_{21}\underline{P}_{11}^{-1})\underline{P}_{11} = -\underline{P}_{11}\underline{\Lambda} \quad . \tag{30}$$

Eq. (30) displays the fact that the poles of the optimal, closed-loop regulator system are the negative-real-part eigenvalues of the Riccati system matrix $\underline{H}$. In addition, the eigenvectors for the closed-loop system matrix are obtained from the eigenvectors for $\underline{H}$. There is now complete justification for requiring $\underline{H}$ to have only nonzero-real-part eigenvalues: not only is the limit argument leading to Eq. (22) plausible, but also the optimal closed-loop system is stable.

Curiously enough, the result of Eq. (30) has been quoted and used in the control theory literature although, to the best of this author's knowledge, no proof of the fact has ever before been published. For Example, Tyler [7] bases his whole theoretical development upon the fact stated in Eq. (30) but provides neither a proof nor a reference to a published proof; Letov [8] also states the result without proof or reference.

USE OF RESULTS IN SYSTEM DESIGN AND ANALYSIS BY COMPUTER [9] [10]

A FORTRAN program has been written for the CDC-3200 in which the eigenvalues and eigenvectors of $\underline{H}$ are found and used to construct the optimal system. In addition, because of Eq. (30), the transient and frequency responses of the optimal closed-loop system are readily computed [5]. The eigenvalues of $\underline{H}$ are found by the QR algorithm [11]; its eigenvectors by Gaussian elimination with partial pivoting. All computations are performed in single precision (36 bits). In comparison with the conventional integration approach, the eigenvalue-eigenvector technique of this paper yields greater accuracy and more information. And in the somewhat unusual problems – specifically the presence of a closed-loop pole with large magnitude real part – the eigenvalue-eigenvector approach gives a solution when the integration method falls apart due to roundoff error accumulation.

REFERENCES

[1] Kalman, R.E., "Contributions to the Theory of Optimal Control," Bol. Soc. Mat. Mexicana, Second Series, vol. 5, pp. 102-119 (1960).

[2] Kalman, R.E., T.S. Englar, and R.S. Bucy, Fundamental Study of Adaptive Control Systems, Tech. Report No. ASD-TR-61-27, vol. I, Wright-Patterson AFB, Ohio (April 1962).

[3] Diliberto, S.P., "On Stability of Linear Mechanical Systems," Trans. International Symposium on Nonlinear Oscillations, vol. I, pp. 189-203 (Kiev, 1963).

[4] Marcus, M., "Matrices in Linear Mechanical Systems," Canadian Math. Bull., vol. 5, no. 3, pp. 253-257 (September 1962).

[5] O'Donnell, J., *Stationary Linear Optimal Control Systems: An Algebraic Approach to Synthesis and Analysis*, RN No. 640, Sylvania Applied Research Laboratory, Waltham, Mass. (27 April 1966).

[6] Potter, J.E., "Matrix Quadratic Solutions," *SIAM Journal of Applied Math.*, vol. 14, no. 3, pp. 496-501 (May 1966).

[7] Tyler, J.S., and F.B. Tuteur, "The Use of a Quadratic Performance Index to Design Multivariable Control Systems," *IEEE Trans. on Automatic Control*, vol. AC-11, no. 1, pp. 84-92 (January 1966).

[8] Letov, A.M., "Analytical Controller Design I," *Automation and Remote Control*, vol. 21, no. 4, (English Translation) pp. 303-306 (Nov. 1960).

[9] Kalman, R.E., and T.S. Englar, *A User's Manual for the Automatic Synthesis Program (Prog. C)*, NASA Report CR-475, Prepared by Martin-Marietta Corp., Baltimore, Maryland (June 1966).

[10] O'Donnell, J., and J. Osborn, *Computer Design of Time-Invariant Optimal Control Systems*, RN No. 655, Sylvania Applied Research Laboratory, Waltham, Mass. (1 July 1966).

[11] Wilkinson, J.H., *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

APPENDIX [3],[4]

Hamiltonian and symplectic matrices are defined and some of their properties are demonstrated in this Appendix.

With $\underline{I}$ and $\underline{0}$ denoting, respectively the identity and null matrices, define the matrix

$$\underline{J} \equiv \left[\begin{array}{c|c} \underline{0} & -\underline{I} \\ \hline \underline{I} & \underline{0} \end{array}\right] .$$

It is easy to demonstrate the following properties of $\underline{J}$: $\underline{J}^2 = -\underline{I}$, $\underline{J}^T = -J$, $\underline{J}^{-1} = -\underline{J}$.

Definition: An even-dimensional matrix $\underline{H}$ is said to be Hamiltonian if $\underline{H} = \underline{J}\,\underline{H}^T J$.

Lemma: For each eigenvalue λ_i of a Hamiltonian matrix, $-\lambda_i$ is also an eigenvalue. In addition, if $\underline{\xi}_i$ and $\underline{\zeta}_i$ are right eigenvectors for λ_i and $-\lambda_i$ respectively, then $-\underline{J}\,\underline{\zeta}_i$ and $\underline{J}\,\underline{\xi}_i$ are the normalized left eigenvectors for λ_i and $-\lambda_i$ respectively.

Proof: Substitution of $\underline{H} = \underline{J}\,\underline{H}^T\underline{J}$ into $\underline{H}\,\underline{\xi}_i = \lambda_i\underline{\xi}_i$ results in $\underline{H}^T\underline{J}\,\underline{\xi}_i = -\lambda_i\underline{J}\,\underline{\xi}_i$, or $(\underline{J}\,\underline{\xi}_i)^T\underline{H} = -\lambda_i(\underline{J}\,\underline{\xi}_i)^T$. Thus $-\lambda_i$ is an eigenvalue of $\underline{H}$ with left eigenvector $\underline{J}\,\underline{\xi}_i$. Let $\underline{\zeta}_i$ be the right eigenvector corresponding to $-\lambda_i$: i.e., $\underline{H}\,\underline{\zeta}_i = -\lambda_i\zeta_i$. Again substituting $\underline{J}\,\underline{H}^T\underline{J}$ for $\underline{H}$, there is obtained $(\underline{J}\,\underline{\zeta}_i)^T\underline{H} = \lambda_i(\underline{J}\,\underline{\zeta}_i)^T$. However, in order to construct similarity transforming matrices for $\underline{H}$ it is necessary to normalize the biorthogonal set formed by the left and right eigenvectors. $\underline{J}\,\underline{\zeta}_i$ is inadmissible as the left eigenvector corresponding to λ_i since the two normalizing relationships

$$(\underline{J}\,\underline{\zeta}_i)^T\underline{\xi}_i = 1 \quad , \qquad (\underline{J}\,\underline{\xi}_i)^T\underline{\zeta}_i = 1$$

are contradictory. Since $-\underline{J}\,\underline{\zeta}_i$ could as well be a left eigenvector for $\underline{H}$ and since $(-\underline{J}\,\underline{\zeta}_i)^T\underline{\xi}_i = (\underline{J}\,\underline{\xi}_i)^T\underline{\zeta}_i = 1$ is a consistent normalizing relationship, $-\underline{J}\,\underline{\zeta}_i$ will be chosen as the normalized left eigenvector for λ_i.

From a practical standpoint, there is no loss of generality in assuming $\underline{H}$ to have nonzero,[*] distinct eigenvalues. Thus, there exists a nonsingular matrix $\underline{P}$ such that

$$\underline{P}^{-1}\underline{H}\,\underline{P} = \underline{\Lambda}_1 \quad (a) , \qquad \text{where} \quad \underline{\Lambda}_1 = \left[\begin{array}{c|c} \underline{\Lambda} & \underline{0} \\ \hline \underline{0} & -\underline{\Lambda} \end{array}\right] \quad (b) \quad . \qquad (i)$$

$$\underline{\Lambda} = \mathrm{diag}\left[\lambda_1, \lambda_2, \cdots, \lambda_n\right] \quad , \qquad \lambda_i \text{ real or complex} \quad . \qquad (ii)$$

In particular, the matrix $\underline{P}$ is given by

$$\underline{P} = \left[\underline{\xi}_1, \cdots, \underline{\xi}_n, \underline{\zeta}_1, \cdots, \underline{\zeta}_n\right] \quad . \qquad (iii)$$

Because the right and left eigenvectors form a bi-orthogonal and normalized set, $\underline{P}^{-1}$ may be constructed from the left eigenvectors. Consequently, using the normalized left eigenvectors, $\underline{P}^{-1}$ becomes[**]

$$\underline{P}^{-1} = \left[-\underline{J}\,\underline{\zeta}_1, \cdots, -\underline{J}\,\underline{\zeta}_n, \underline{J}\,\underline{\xi}_1, \cdots, \underline{J}\,\underline{\xi}_n\right]^T \quad . \qquad (iv)$$

By comparing (iv) and (iii) it would seem that there should exist some transformation from $\underline{P}$ to $\underline{P}^{-1}$. Such a transformation does exist and it will now be exhibited. Let

$$\underline{P}_1 = \left[\underline{\xi}_1, \cdots, \underline{\xi}_n\right]_{2n \times n} , \qquad \underline{P}_2 = \left[\underline{\zeta}_1, \cdots, \underline{\zeta}_n\right]_{2n \times n} \quad . \qquad (v)$$

Then

$$\underline{P} = \left[\underline{P}_1 \,\middle|\, \underline{P}_2\right] \quad (a) , \qquad \underline{P}^{-1} = \left[\begin{array}{c} -(\underline{J}\,\underline{P}_2)^T \\ \hline (\underline{J}\,\underline{P}_1)^T \end{array}\right] = \left[\begin{array}{c} \underline{P}_2^T\underline{J} \\ \hline -\underline{P}_1^T\underline{J} \end{array}\right] \quad (b) \quad . \qquad (vi)$$

Now form the product

$$-\underline{J}\,\underline{P}^T\underline{J} = \left[\begin{array}{c|c} \underline{0} & \underline{I} \\ \hline -\underline{I} & \underline{0} \end{array}\right] \left[\begin{array}{c} \underline{P}_1^T\underline{J} \\ \hline \underline{P}_2^T\underline{J} \end{array}\right] = \left[\begin{array}{c} \underline{P}_2^T\underline{J} \\ \hline -\underline{P}_1^T\underline{J} \end{array}\right] \quad . \qquad (vii)$$

Thus $$-\underline{J}\,\underline{P}^T\underline{J} = \underline{P}^{-1} \qquad \text{or} \qquad \underline{P}^T\underline{J}\,\underline{P} = \underline{J} \quad . \qquad (viii)$$

---

[*] Since the eigenvalues of the optimal closed loop system are the negative-real-part eigenvalues of $\underline{H}$, the closed-loop system is stable only if $\underline{H}$ has nonzero eigenvalues.

[**] Recall that left and right eigenvectors are orthogonal so that

$$(-\mathrm{J}\,\underline{\zeta}_i)^T\underline{\xi}_j = 0 \quad , \qquad (\underline{J}\,\underline{\xi}_i)^T\underline{\zeta}_j = 0 \quad , \qquad i \neq j \quad .$$

Defintion: An even-dimensional matrix $\underline{S}$ is said to be symplectic if $\underline{S}^T\underline{J}\ \underline{S} = \underline{J}$. The following theorem has thus been proved, and follows directly from Eqs. (ia), (ib), and (viii).

Theorem: If a Hamiltonian matrix $\underline{H}$ has nonzero, distinct eigenvalues, there there exists a symplectic matrix $\underline{P}$ such that

$$\underline{P}^{-1}\underline{H}\ \underline{P} = \left[\begin{array}{c|c} \underline{\Lambda} & \underline{0} \\ \hline \underline{0} & -\underline{\Lambda} \end{array}\right] ,$$

where

$$\underline{\Lambda} = \mathrm{diag}\left[\lambda_1, \lambda_2, \cdots, \lambda_n\right] .$$

In words, $\underline{H}$ is symplectically similar to a diagonal matrix.

It is believed that this is the first publication of the symplectic similarity property of Hamiltonian matrices.

DISCRETE LINEAR SYSTEM WITH CONVEX COST FUNCTIONAL*

DONG HAK CHYUNG
Department of Electrical Engineering
University of South Carolina
Columbia, South Carolina

ABSTRACT

This paper considers the optimal control problem of linear discrete systems with convex cost functional. Necessary and sufficient conditions for an optimal control are derived, and the existence of an optimal control is proved.

The optimal control problem of discrete linear systems with quadratic cost functional has been studied in [1]. The purpose of this paper is to extend the results obtained in [1] to systems with more general convex cost functionals. Most of the results in this paper are direct generalizations of the results in [1], and hence only rough sketches are given except when they are significantly different. For more details the reader is referred to [1].

Consider the system

$$x(k+1) = A(k)x(k) + B(k)u(k)$$

where $x(k) \varepsilon R^n$ is an n-dimensional column vector, $A(k)$ is an $n \times n$ matrix with $\det A(k) \neq 0$ for all k, $B(k)$ is an $n \times m$ matrix and $u(k) \varepsilon R^m$ is an m-dimensional column vector. The optimal control problem is to find a control sequence $\{u(k)\} = \{u(0), u(1), \ldots, u(N-1)\}$ from a given compact convex control restraint set $\Omega \subset R^m$ such that the corresponding response $x(k)$ is steered from a given initial point $x(0) = x_0$ to a given closed convex target set $G \subset R^n$ at $k = N$, and at the same time the cost functional

$$C(u) = g(x(N)) + \sum_{k=0}^{N-1} \{f(k,x(k)) + h(k,u(k))\}$$

is minimized. Here N is a fixed number, $g(x)$ is a C^1 convex real function of $x \varepsilon R^n$, $f(k,x)$ is a nonnegative real C^1 convex function of $x \varepsilon R^n$ for all k and $h(k,u)$ is a nonnegative convex continuous real function of $u \varepsilon R^m$ for all k.

Let

$$x^\circ(k) = \sum_{j=0}^{k-1} \{f(j,x(j)) + h(j)u(j))\}, \quad x^\circ(0) = 0$$

and define a new (n+1)-column vector $\hat{x}(k) \varepsilon R^{n+1}$ by $\hat{x}(k) = (x^\circ(k), x(k))$. Denote the response of the new n+1 dimensional system corresponding to a control $\{u(k)\}$ by $\hat{x}_u(k) = (x^\circ{}_u(k), x_u(k))$. Then

$$x_u(k) = W(k)x_0 + W(k) \sum_{j=0}^{k-1} W^{-1}(j+1)B(j)u(j)$$

and

---

*Research supported in part by Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Grant No. AF-AFOSR-571-66 while the author was at the center for Control Sciences, University of Minnesota, Minneapolis.

$$x^\circ_u(k) = \sum_{j=0}^{k-1} \{f(j,x_u(j)) + h(j,u(j))\}, \quad x^\circ_u(0) = 0$$

where $W(k)$ is the $n\times n$ matrix solution of

$$W(k+1) = A(k)W(k),\ W(0) = I, \text{ the identity matrix.}$$

Define the sets of attainability $\hat{K}\subset R^{n+1}$ and $K\subset R^n$ to be the sets of the response endpoints $\hat{x}_u(N)$ and $x_u(N)$ for all admissible controls respectively, that is,

$$\hat{K} = \{\hat{x}_u(N) \mid \hat{x}_u(0) = \hat{x}_0,\ \{u(k)\} \subset \Omega\}$$

$$K = \{x_u(N) \mid x_u(0) = x_0,\ \{u(k)\}\subset\Omega\}$$

A control $\{u(k)\}$ is called admissible if $u(k)\in\Omega$ for all k.

Since $f(x,k) \geq 0$ and $h(u,k) \geq 0$, $x^\circ_u(k)\geq 0$ for all $\{u(k)\}$ and so $\hat{K}$ is in the upper half space $x^\circ \geq 0$ of R^{n+1}. From the definitions it is obvious that K is the orthogonal projection of $\hat{K}$ on the hyperplane $x^\circ = 0$ in R^{n+1}. We assume that $\hat{K}$ has a nonempty interior in R^{n+1}, otherwise one can apply the same analysis in the linear manifold which is spanned by $\hat{K}$. Also we assume that the basic linear system is irrational of degree N (see [2]).

Now the optimal control problem can be rephrased as follows: Find a point $\hat{x}^* = (x^{*\circ},x^*)$ in $\hat{K}\cap\hat{G}$ such that $x^{*\circ} = \inf x^\circ$ for all $\hat{x} = (x^\circ,x)$ in $\hat{K}\cap\hat{G}$, here $\hat{G} = R^1\times G\subset R^{n+1}$. The optimal control is a control which steers the response to x^*.

Theorem 1. The set of attainability $\hat{K}$ is compact in R^{n+1}.
Sketch of Proof: $\hat{K}$ is a continuous mapping of the compact set $\Omega\times\Omega\times\cdots\times\Omega\subset R^{Nm}$ into R^{n+1}. Therefore $\hat{K}$ is compact.

It was shown in [2] that K is compact and convex in R^n. Define the lower boundary (with respect to x°-axis) $\partial\hat{K}$ of $\hat{K}$ over the compact convex set K to be the set of all lower limit points of the compact set $\hat{K}$, that is,

$$\partial\hat{K} = \{\hat{x} = (x^\circ,x) \mid x\in K,\ x^\circ = \inf_{(y^\circ,x)\in\hat{K}} \{y^\circ\}\}$$

Theorem 2. The lower boundary $\partial\hat{K}$ of $\hat{K}$ is a convex hypersurface in R^{n+1} over the compact convex set K.
Sketch of Proof: Let $\hat{x}_1 = (x^\circ_1,x_1)$ and $\hat{x}_2 = (x^\circ_2,x_2)$ be any two points on $\partial\hat{K}$ corresponding to controls $\{u_1(k)\}$ and $\{u_2(k)\}$ respectively. To prove the theorem it is sufficient to show that for any point $\hat{x}=\lambda\hat{x}_1+(1-\lambda)\hat{x}_2$, $0<\lambda<1$, there exists a point $\hat{x}^* = (x^{\circ *},x^*)$ in $\hat{K}$ such that $x^{\circ *}\leq \lambda x^\circ_1+(1-\lambda)x^\circ_2=x^\circ$, $x^* = \lambda x_1 + (1-\lambda)x_2 = x$, for then there exists a point $\hat{y} = (y^\circ,y)$ on $\partial\hat{K}$ such that $y^\circ\leq x^\circ$, $y = x$.

Define a control $u^*(k) = \lambda u_1(k) + (1-\lambda)u_2(k)$. Then $\{u^*(k)\}\subset\Omega$ for Ω is convex, and so the corresponding response $\hat{x}^*(N)$ is in $\hat{K}$. It is easy to show that $x^*(N) = \lambda x_1(N) + (1-\lambda)x_2(N) = x$. Also, from the convexity of $f(k,x)$ and $h(k,u)$, it can be shown that $x^{\circ *}(N)\leq\lambda x^\circ_1(N) + (1-\lambda)x^\circ_2(N) = x^\circ$. Therefore $\partial\hat{K}$ is a convex hypersurface.

It is shown later that an optimal control steers the response to the lower boundary $\partial\hat{K}$. Therefore the lower boundary points are very important. The following maximum principle characterizes those controls which steer their responses to the boundary $\partial\hat{K}$.

Theorem 3. An admissible control $\{u^*(k)\}\subset\Omega$ steers its response $\hat{x}^*(k)$ to the boundary $\partial\hat{K}$ if and only if there exists an (n+1)-column vector adjoint solution $\hat{q}(k) = (q_0, q(k))\in R^{n+1}$ of the adjoint equation

$$q_0 \le 0, \text{ constant}$$

$$q(k) = q(k+1)A(k) + q_0 \frac{\partial f}{\partial x}(k,x^*(k))$$

such that

$$q_0 h(k,u^*(k)) + q(k+1)B(k)u^*(k) = \max_{u\in\Omega} q_0 h(k,u) + q(k+1)B(k)u$$

for all k = 0, 1, ..., N-1.

Proof: Suppose a control $\{u^*(k)\}$ with response $\hat{x}^*(k)$ satisfies the maximum principle. Let $\{u(k)\}$ be an arbitrary admissible control with response $\hat{x}(k)$. Then compute, remembering $x^\circ(0) = x^{*\circ}(0) = 0$.

$$\begin{aligned}
\hat{q}(N)\hat{x}(N) - \hat{q}(0)\hat{x}(0) &= q_0 x^\circ(N) + q(N)x(N) - q_0 x^\circ(0) - q(0)x(0) \\
&= q_0 x^\circ(N) + \sum_{k=0}^{N-1} \{q(k+1)x(k+1) - q(k)x(k)\} \\
&= q_0 \sum_{k=0}^{N-1} \{f(k,x(k)) + h(k,u(k))\} \\
&\quad + \sum_{k=0}^{N-1} \{q(k+1)(A(k)x(k) + B(k)u(k)) \\
&\quad - (q(k+1)A(k) + q_0 \frac{\partial f}{\partial x}(k,x^*(k)))x(k)\} \\
&= q_0 \sum_{k=0}^{N-1} \{f(k,x(k)) - q_0 \frac{\partial f}{\partial x}(k,x(k))\} \\
&\quad + \sum_{k=0}^{N-1} \{q_0 h(k,u(k)) + q(k+1)B(k)u(k)\}
\end{aligned}$$

For $\{u^*(k)\}$ with response $\hat{x}^*(k)$,

$$\begin{aligned}
\hat{q}(N)\hat{x}^*(N) - \hat{q}(0)\hat{x}^*(0) &= q_0 \sum_{k=0}^{N-1} \{f(k,x^*(k)) - q_0 \frac{\partial f}{\partial x}(k,x^*(k))\} \\
&\quad + \sum_{k=0}^{N-1} \{q_0 h(k,u^*(k)) + q(k+1)B(k)u^*(k)\}
\end{aligned}$$

From the maximum principle

$$q_0 h(k,u^*(k)) + q(k+1)B(k)u^*(k) \ge q_0 h(k,u(k)) + q(k+1)B(k)u(k)$$

Also from the convexity of f(k,x)

$$f(k,x(k)) - f(k,x^*(k)) - \frac{\partial f}{\partial x}(k,x^*(k))(x(k) - x^*(k)) \ge 0$$

Since $q_0 \le 0$

$$q_0 \{f(k,x^*(k)) - \frac{\partial f}{\partial x}(k,x^*(k))x^*(k)\} \ge q_0 \{f(k,x(k)) - \frac{\partial f}{\partial x}(k,x^*(k))x(k)\}$$

Thus, by comparing the two equations

$$\hat{q}(N)\hat{x}^*(N) - \hat{q}(0)\hat{x}^*(0) \ge \hat{q}(N)\hat{x}(N) - \hat{q}(0)\hat{x}(0)$$

Now

$$\hat{q}(0)\hat{x}^*(0) = q(0)x_0 = \hat{q}(0)\hat{x}(0).$$

Therefore

$$\hat{q}(N)\hat{x}^*(N) \geq \hat{q}(N)\hat{x}(N)$$

for any $\hat{x}(N)$ in $\hat{K}$. If $q_0 < 0$, then since the lower boundary $\partial\hat{K}$ of $\hat{K}$ is a convex hypersurface over K, $\hat{x}^*(N)$ must be on $\partial\hat{K}$. If $q_0 = 0$ then $x^*(N)$ must be a boundary point of K. But then, since the basic system is irrational, $\{u^*(k)\}$ is the only admissible control which steers the response to $x^*(N)$. Therefore $\hat{x}^*(N)$ is the only point in $\hat{K}$ with $x = x^*(N)$. Thus $\hat{x}^*(N)$ is a lower boundary point of $\hat{K}$.

Conversely, suppose a control $\{u^*(k)\}$ steers the response $\hat{x}^*(k)$ to the lower boundary, that is, $\hat{x}^*(N)\varepsilon\partial\hat{K}$. Then, since $\partial\hat{K}$ is a convex hypersurface over K, there exists a tangent hyperplane to the set $\hat{K}$ at $x^*(N)\varepsilon\partial\hat{K}$ in R^{n+1}. Let $\hat{q} = (q^\circ, q)$ be a normal vector to the tangent hyperplane directing into the half-space which is defined by the tangent hyperplane and does not contain $\hat{K}$, that is, $\hat{q}$ is an outward normal vector to $\hat{K}$ at $\hat{x}^*(N)$. Clearly $q^\circ \leq 0$. If $q^\circ = 0$, then $x^*(N)$ is a boundary point K and q is an outward normal to K at $x^*(N)$. Therefore if $\hat{q}(k) = (q_0, q(k))$ is a solution of the adjoint equation with the endpoint condition $\hat{q}(N) = \hat{q}$, that is, $q_0 = q^\circ$, $q(N) = q$, then $\{u^*(K)\}$ must satisfy the maximum principle. (See [2]). If $q^\circ < 0$ then let $\hat{q}(k)$ be the adjoint solution with the endpoint condition $\hat{q}(N) = \hat{q}$. Now suppose $\{u^*(k)\}$ does not satisfy the maximum principle with $\hat{q}(k)$. Then for at least one k, say k_0, there exists a nonzero constant $\delta > 0$ such that

$$q_0 h(k_0, u^*(k_0)) + q(k_0+1)B(k_0)u^*(k_0) + \delta \leq \max_{u\varepsilon\Omega} q_0 h(k_0, u) + q(k_0+1)B(k_0)u$$

Since $h(k,u)$ is continuous in u, for any ε, $0<\varepsilon<\delta$, there exists a $\bar{u}$ in Ω such that

$$q_0 h(k_0, u^*(k_0)) + q(k_0+1)B(k_0)u^*(k_0) + \varepsilon = q_0 h(k_0, \bar{u}) + q(k_0+1)B(k_0)\bar{u}$$

Define a new control $\{u_\varepsilon(k)\}$

$$u_\varepsilon(k) = \begin{cases} u^*(k) & \text{if } k \neq k_0 \\ \bar{u} & \text{if } k = k_0 \end{cases}$$

and let $\hat{x}_\varepsilon(k)$ be the corresponding response. Then $\hat{x}_\varepsilon(N)\varepsilon\hat{K}$ for $\{u_\varepsilon(k)\} \subset \Omega$. Now, compute as before,

$$\begin{aligned} \hat{q}(N)\hat{x}^*(N) - \hat{q}(N)\hat{x}_\varepsilon(N) = q_0 \sum_{k=0}^{N-1} & \{f(k,x^*(k)) - f(k,x_\varepsilon(k)) \\ & - \frac{\partial f}{\partial x}(k,x^*(k))(x^*(k) - x_\varepsilon(k))\} + \sum_{k=0}^{N-1} \{q_0 h(k,u^*(k)) \\ & + q(k+1)B(k)u^*(k) - q_0 h(k,u\ (k)) - q(k+1)B(k)u(k)\} \end{aligned}$$

$$= q_0 \sum_{k=0}^{N-1} \{f(k,x^*(k)) - f(k,x_\varepsilon(k)) - \frac{\partial f}{\partial x}(k,x^*(k)) - (x^*(k) - x_\varepsilon(k))\} - \varepsilon$$

It is easy to show that $x_\varepsilon(k)$ is continuous in ε for each k. Furthermore $f(k,x)$ is a C^1 function in x and $x_{\varepsilon=0}(k) = x^*(k)$ by the definition. Therefore, for sufficiently small $\varepsilon<\delta$, there exists a constant c_1 such that

$$|f(k,x^*(k)) - f(k,x_\varepsilon(k)) - \frac{\partial f}{\partial x}(k,x^*(k))(x^*(k)-x_\varepsilon(k))| < c_1\varepsilon^2$$

for all k. Then

$$\hat{q}(N)\hat{x}^*(N) - \hat{q}(N)\hat{x}_\varepsilon(N) < c_2\varepsilon^2 - \varepsilon$$

for some constant c_2, and so

$$\hat{q}(N)\hat{x}^*(N) < \hat{q}(N)\hat{x}_\varepsilon(N)$$

for sufficiently small $\varepsilon > 0$. However this is impossible, for $\hat{q}(N) = \hat{q}$ is an outward normal vector to $\hat{K}$ at the lower boundary point $\hat{x}^*(N) \varepsilon \partial\hat{K}$, $\partial\hat{K}$ is a convex hypersurface and $\hat{x}_\varepsilon(N) \varepsilon \hat{K}$. Therefore $\{u^*(k)\}$ must satisfy the maximum principle with respect to the adjoint solution $\hat{q}(k)$.

Suppose the target set G is the whole space R^n (the free endpoint problem), and consider a cost functional $C(u) = g(x(N)) + x^\circ(N)$. Then, since $g(x)$ is continuous in x, $C(u)$ is a continuous function of $\hat{x}$. Therefore there exists a minimum of $C(u)$ on the compact set $\hat{K}$ in R^{n+1}, and hence there exists an optimal control. Furthermore, since $g(x)$ is convex and $C(u)$ is strictly increasing in x° for fixed x, it can be shown that an optimum point $\hat{x}^*$, at which $C(u)$ assumes its minimum, must be a lower boundary point of $\hat{K}$, that is, $\hat{x}^* \varepsilon \partial\hat{K}$. Therefore, by Theorem 3, an optimal control $\{u^*(k)\}$ which steers the response $\hat{x}^*(k)$ to $\hat{x}^*$ must satisfy the maximum principle with $\hat{q}(k)$. Here $\hat{q}(k)$ is an adjoint solution such that $\hat{q}(N)$ is outward normal to $\hat{K}$ at $\hat{x}^* \varepsilon \partial\hat{K}$. This last condition is often called transversality condition. Since $g(x)$ is a C^1 convex function, the transversality condition is satisfied if $q_0 = -1$, $q(N) = -\text{grad } g(x^*(N))$. Also, by Theorem 3, if a control $\{u^*(k)\}$ with response $\hat{x}^*(k)$ satisfies the maximum principle with the transversality condition then $\hat{x}^*(k)$ is steered to an optimal point $\hat{x}^*$, and so $\{u^*(k)\}$ is an optimal control. For more details see [1]. The following two theorems summarize the above results.

Theorem 4. Consider the free endpoint point problem, that is, $G = R^n$ with the cost functional $C(u) = g(x(N)) + x^\circ(N)$. Then there exists an optimal control.

Theorem 5. Consider the cost functional $C(u) = g(x(N)) + x^\circ(N)$. Then there exists a solution $x^*(k), q^*(k)$ of the equation

$$x(k+1) = A(k)x(k) + B(k)u^*(k,q)$$

$$q(k) = q(k+1)A(k) + q_0 \frac{\partial f}{\partial x}(k,x(k))$$

with $q(N) = -\text{ grad } g(x(N))$, $q_0 = -1$. $x(0) = x_0$. Here $u^*(k,q)$ is defined by the maximum principle

$$q(k+1)B(k)u^*(k,q) + q_0 h(k,u^*(k,q)) = \max_{u\varepsilon\Omega} q(k+1)B(k)u + q_0 h(k,u)$$

An optimal control is given by $u^*(k) = u^*(k,q^*)$.

Remarks: If the target G is a closed convex subset of R^n instead of the whole space R^n as in Theorems 4 and 5, then the same results can be obtained except that the endpoint condition must be modified to satisfy the transversality condition. See [1] for more details.

References

1. Chyung, D.H., "Discrete Linear Optimal System with Essentially Quadratic Cost Functional," IEEE Trans. on Automatic Control, Vol. AC-11, No.3, July 1966.

2. Lee, E.B., "Recurrence Equations and the Control of Their Evolution," J. of Math. Anal. and Appl., Vol. 7, No. 1, August 1963.

3. Lee, E.B., and Markus, L., Foundation of Optimal Control, John Wiley, New York, 1966.

4. Pontryagin, L.S., et. al., The Mathematical Theory of Optimal Processes, Interscience, New York, 1962.

A SUFFICIENCY CONDITION FOR MINIMUM VARIANCE CONTROL OF MARKOV PROCESSES

M. K. SAIN
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

EXORDIUM

The present investigation has its analytical origin in the work of Wonham [1] and its conceptual origin in the author's continuing attention to minimum variance control problems [2,3,4,5].

For simplicity, the present development adheres to the notation of [1] and begins by recalling some results from that study. Consider a plant of the form

$$\dot{x} = f(t,x,u(t),\alpha(t)), \quad t_0 \leq t \leq t_F , \tag{1}$$

where x and f are n-vectors, u(t) is an m-vector control, and $\alpha(t)$ is a q-vector random process. If the controller has perfect information about the state of the system, then a control law of the form

$$u(t) = k(t,x(t)), \quad t_0 \leq t \leq t_F , \tag{2}$$

may be sought. Equations (1) and (2) define a random process

$$\dot{x} = f(t,x,k(t,x),\alpha(t)), \quad t_0 \leq t \leq t_F , \tag{3}$$

which is assumed to be Markov, with known backward equation

$$\frac{-\partial P(t,x;\ s,E;\ k)}{\partial t} = \mathcal{L}(k)[P(t,x;\ s,E;\ k)]. \tag{4}$$

In (4), $\mathcal{L}(k)$ is a linear operator which acts on x and depends on the control law k; $P(t,x;\ s,E;\ k)$ is the transition probability distribution under k, i.e.,

$$P(t,x;\ s,E;\ k) = \Pr[x(s) \in E \mid x(t) = x;\ k], \quad s \geq t \geq t_0; \tag{5}$$

the terminal condition is given by

$$\begin{aligned} \lim_{t \to s} P(t,x;\ s,E;\ k) &= 1 \ , \ x \in E \ , \\ &= 0 \ , \ x \notin E. \end{aligned} \tag{6}$$

Suppose that it is desired to minimize the expected value of a cost functional

$$\int_{t_0}^{t_F} L(s,x(s),u(s))ds + \lambda(x(t_F)) = J(t_0,x_0,u); \tag{7}$$

by imbedding procedure, define

$$V_M(t,x;\ k) = E\{J(t,x;\ k) \mid x(t) = x\} . \tag{8}$$

Then Wonham has stated the following results.
I. $V_M(t,x;\ k)$ satisfies the partial differential equation

$$-\frac{\partial V_M(t,x;\, k)}{\partial t} = L(t,x,k(t,x)) + \mathcal{L}(k)[V_M(t,x;\, k)] \,, \tag{9}$$

together with the terminal condition

$$V_M(t_F,x;\, k) = \lambda(x) \,. \tag{10}$$

II. If there is a control law k_M^o such that $V_M(t,x;\, k_M^o)$ satisfies (9) and (10), and such that

$$\begin{aligned} &\mathcal{L}(k^o{}_M)[V_M(t,x;\, k_M^o] + L(t,x,k_M^o(t,x)) \\ &\leq \mathcal{L}(k)[V_M(t,x;\, k_M^o] + L(t,x,k(t,x)) \end{aligned} \tag{11}$$

for all t,x in the domain of interest and for all admissible control laws k, then k_M^o is an optimal control law, i.e.,

$$V_M(t,x;\, k_M^o) = V_M^o(t,x) \leq V_M(t,x;\, k) \tag{12}$$

for all t,x in the domain of interest and admissible k.

III. $V_M^o(t,x)$ satisfies the Stochastic Hamilton-Jacobi equation

$$\frac{-\partial V_M^o(t,x)}{\partial t} = \underset{u}{\mathrm{Min}}\; (L(t,x,u) + \mathcal{L}(u)[V_M^o(t,x)]) \tag{13}$$

where the minimization is carried out over admissible u, defined by admissible k.

In the sequel, counterparts of these results are obtained for a class of minimum variance control problems.

PROBLEM DEFINITION AND MAIN RESULTS

Let the plant be described by (1), the controller by (2), and the cost functional by (7). It is assumed that the controlled diffusion process (3) is Markov, with evolution according to (4), (5), and (6). In the present instance, however, the minimum of (8) is not in general required; instead, it is constrained according to

$$V_M(t,x;\, k) = M(t,x) \tag{14}$$

for all t,x of interest, where M(t,x) is given in advance. Clearly,

$$M(t,x) \geq V_M^o(t,x) \,. \tag{15}$$

Under the constraint (14), it is desired to minimize the variance of the cost function, and so follows the definition

$$V_V(t,x;\, k) = E\{J^2(t,x;\, k) \mid x(t) = x,\, V_M(t,x;\, k) = M(t,x)\}. \tag{16}$$

The development begins by establishing Theorem I, which corresponds to (9).

Theorem I. The cost variance (16) associated with the controlled diffusion process (3) satisfies the partial differential equation

$$-\frac{\partial V_V(t,x;\, k)}{\partial t} = 2L(t,x,k(t,x))M(t,x) + \mathcal{L}(k)[V_V(t,x;\, k)] \tag{17}$$

together with the terminal condition

$$V_V(t_F,x;\ k) = \lambda^2(x)\ . \tag{18}$$

Proof: For clarity, assume that

$$P(t,x;\ s,E;\ k) = \int_E p(t,x;\ s,y;\ k)dy; \tag{19}$$

then, interchanging integration and expectation,

$$\begin{aligned} V_V(t,x;\ k) = \int_t^{t_F}\int_t^{t_F} & drds[\int\int p(t,x;\ r,y;\ s,z;\ k)\ \cdot \\ & \cdot\ L(r,y,k(r,y))L(s,z,k(s,z))dydz]\ + \\ & + 2\int_t^{t_F} dr\int\int p(t,x;\ r,y;\ t_F,z;\ k)\lambda(z)L(r,y,k(r,y))dydz \\ & + \int p(t,x;\ t_F,y;\ k)\lambda^2(y)dy, \end{aligned} \tag{20}$$

where

$$\begin{aligned} & p(t,x;\ r,y;\ s,z;\ k)\ dydz = \\ & = Pr[y<x(r)\leq y+dy, z<x(s)\leq z+dz\ |\ x(t) = x;\ k],\ r \geq t,\ s \geq t \\ & = Pr[z<x(s)\leq z+dz\ |\ x(r) = y,\ x(t) = x;\ k]\ \cdot \\ & \cdot\ Pr[y<x(r)\leq y+dy\ |\ x(t) = x;\ k],\ s \geq r \geq t. \end{aligned} \tag{21}$$

Because (3) is assumed to define a Markov process, it follows that

$$\begin{aligned} & p(t,x;\ r,y;\ s,z;\ k)dydz = \\ & = p(r,y;\ s,z;\ k)dzp(t,x;\ r,y;\ k)dy,\ s \geq r \geq t. \end{aligned} \tag{22}$$

A similar expression holds for r ≥ s ≥ t. An important point to observe is that (22) implies

$$-\frac{\partial p(t,x;\ r,y;\ s,z;\ k)}{\partial t} = \mathcal{L}(k)[p(t,x;\ r,y;\ s,z;\ k]. \tag{23}$$

Further since

$$\lim_{s\to t} p(t,x;\ s,y;\ k) = \delta(y-x)\ , \tag{24}$$

then (22) gives the result

$$\lim_{r\to t} p(t,x;\ r,y;\ s,z;\ k)dydz = p(t,y;\ s,z;\ k)dz\delta(y-x)dy.\ \ s \geq t. \tag{25}$$

With the aid of (22), (23), (24), and (25), the formal derivation of (17) and (18) is straightforward. The second step, corresponding to (11), develops an optimality relation.

Theorem II. Suppose that there exists a control law k_V^o such that $V_V(t,x;\ k_V^o)$ satisfies (17) and (18), and such that

$$\begin{aligned}&\mathcal{L}(k_V^o)[V_V(t,x;\ k_V^o)] + 2L(t,x,k_V^o(t,x))M(t,x)\\ &\le \mathcal{L}(k)[V_V(t,x;\ k_V^o)] + 2L(t,x,k(t,x))M(t,x)\end{aligned} \tag{26}$$

for all t,x in the domain of interest and for all admissible control laws k, then k_V^o is an optimal control law, i.e.,

$$V_V(t,x;\ k_V^o) = V_V^o(t,x) \le V_V(t,x;\ k) \tag{27}$$

for all t,x in the domain of interest and admissible k.

Proof: In order to establish (26), it is convenient to make use of an existing result [1, p.122].

Lemma (Wonham). Let W(s,y) be an arbitrary smooth function defined for $t_0 \le s \le t_F$ and y an n-vector. Then

$$\begin{aligned}&\int p(t,x;\ t_F,y)W(t_F,y)dt = W(t,x) +\\ &+ \int_t^{t_F} ds\int p(t,x;\ s,y)\left[\frac{\partial W(s,y)}{\partial s} + \mathcal{L}[W(s,y)]\right]dy.\end{aligned} \tag{28}$$

To both members of (26), add $\dfrac{\partial V_V^o(t,x)}{\partial t}$, so that

$$0 \le \frac{\partial V_V^o(t,x)}{\partial t} + \mathcal{L}(k)[V_V(t,x;\ k_V^o)] + 2L(t,x,k(t,x))M(t,x) \tag{29}$$

for all t,x of interest. Therefore,

$$\begin{aligned}0 \le &\int_t^{t_F} ds\int p(t,x;\ s,y;\ k)\left[\frac{\partial V_V^o(s,y)}{\partial s} + \mathcal{L}(k)[V_V^o(s,y)] +\right.\\ &\left.+ 2L(s,y,k(s,y))M(s,y)\right]dy.\end{aligned} \tag{30}$$

Application of (28) to (30) yields

$$\begin{aligned}0 \le &\int p(t,x;\ t_F,y;\ k)V_V^o(t_F,y)dy - V_V^o(t,x)\\ &+ 2\int_t^{t_F} ds\int p(t,x;\ s,y;\ k)L(s,y,k(s,y))M(s,y)dy.\end{aligned} \tag{31}$$

Equations (14), (18), and (20), when combined with (31), imply

$$\begin{aligned}V_V^o(t,x) \le &\int p(t,x;\ t_F,y;\ k)\lambda^2(y)dy\\ &+ 2\int_t^{t_F} ds\left[\int\int p(t,x;\ s,y;\ k)dy\,p(s,y;\ t_F,z;\ k)dz\ \cdot\right.\\ &\left.\cdot\lambda(z)L(s,y,k(s,y))\right]\\ &+ 2\int_t^{t_F} ds\left[\int p(t,x;\ s,y;\ k)L(s,y,k(s,y))\ \cdot\right.\\ &\left.\cdot\int_s^{t_F} dr\int p(s,y;\ r,z;\ k)L(r,z,k(r,z))dzdy\right]\end{aligned} \tag{32}$$

which yields the final result (27). With the assistance of (26), a Stochastic Hamilton-Jacobi Equation can be established.

Theorem III. Let the class of admissible control laws k determine the set of admissible values of the control function u; then $V_V^o(t,x)$ satisfies a Stochastic Hamilton-Jacobi equation.

$$-\frac{\partial V_V^o(t,x)}{\partial t} = \min_u \left[2M(t,x)L(t,x,u) + \mathcal{L}(u)[V_V^o(t,x)] \right] \tag{33}$$

where the minimization is carried out over the set of admissible u which satisfy the constraint (14).

Proof: Theorem III follows immediately from Theorem I and Theorem II.

Remark I: Two essential differences distinguish the minimum variance results (17), (18), (26), (27), and (33) from the minimum mean results (9), (10), (11), (12) and (13): the presence of a factor 2M(t,x) multiplying L(t,x,u) and the terminal condition $\lambda^2(x)$ in place of $\lambda(x)$. In other respects, however, the form of the equations for variance and mean problems are very much the same, particularly in regard to the space operator $\mathcal{L}$.

Remark II. An effective approach to the incorporation of constraint (14) is to make use of (9), with the added restriction that M(t,x) possess the required derivatives. In such a case,

$$-\frac{\partial M(t,x)}{\partial t} = L(t,x,k) + \mathcal{L}(k)[M(t,x)], \tag{34}$$

where it is necessary that

$$M(t_F,x) = \lambda(x). \tag{35}$$

Because (34) is an identity for all t,x of interest, (26) can be written

$$\begin{aligned} &\mathcal{L}(k_V^o)[V_V(t,x;\, k_V^o)] - 2M(t,x)\, \mathcal{L}(k_V^o)[M(t,x)] \\ &\le \mathcal{L}(k)[V_V(t,x;\, k_V^o)] - 2M(t,x)\, \mathcal{L}(k)[M(t,x)] \end{aligned} \tag{36}$$

which allows the corresponding revision of (33), namely

$$\begin{aligned} \frac{-\partial V_V^o(t,x)}{\partial t} &= \min_u \left[\mathcal{L}(u)[V_V^o(t,x)] - 2M(t,x)\,\mathcal{L}(u)[M(t,x)] \right] \\ &\quad - 2M(t,x)\,\frac{\partial M(t,x)}{\partial t}, \end{aligned} \tag{37}$$

where again u is chosen from admissible controls satisfying (34).

EXAMPLE

In order to clarify the above ideas, consider a very simple example, in which

$$f(t,x,u(t),\alpha(t)) = u(t) + \alpha(t),\quad t_0 \le t \le t_F \tag{38}$$

$$L(t,x,u) = \tfrac{1}{2}x^2 + \tfrac{1}{2}u^2,\quad \lambda(x) = 0, \tag{39}$$

and $\alpha(t)$ is gaussian white noise with

$$E\{\alpha(t)\alpha(s)\} = b\delta(s-t). \tag{40}$$

For this case we have

$$\mathcal{L}(k) = \frac{1}{2}b\,\frac{\partial^2}{\partial x^2} + k\,\frac{\partial}{\partial x}\,. \tag{41}$$

Assume that $M(t,x)$ is of the form

$$M(t,x) = m(t)x^2 + b\int_t^{t_F} m(s)ds,\ m(t_F) = 0; \tag{42}$$

then from (34),

$$0 = \dot{m}x^2 + \frac{1}{2}(x^2+u^2) + 2mxu, \tag{43}$$

so that there are two choices for control

$$u_{1,2} = (-2m \pm \sqrt{4m^2 - (2\dot{m}+1)}\ )x = k_{1,2}(t)x. \tag{44}$$

Remark III. The discriminant in (44) vanishes when $m(t)$ is chosen to satisfy (13), so that there is no choice left in (37). Thus a relaxation of the minimum average cost requirement allows the opportunity to reduce cost variance.

Then (37) becomes

$$\begin{aligned}\frac{-\partial V_V^o(t,x)}{\partial t} = \min_{k_{1,2}} \{\frac{1}{2}b\,&\frac{\partial^2 V_V^o(t,x)}{\partial x^2} + k_i x\,\frac{\partial V_V^o(t,x)}{\partial x}\\ &- 2[mx^2 + b\int_t^{t_F} m(s)ds][mb + 2k_i mx^2]\}\\ &- 2[mx^2 + b\int_t^{t_F} m(s)ds][\dot{m}x^2 - b\,m],\end{aligned} \tag{45}$$

so that the optimal controller is given by

$$u_V^o = -\ [+2m + (\mathrm{sgn}\{x\,\frac{\partial V_V^o(t,x)}{\partial x} - 4mx^2M(t,x)\})\cdot\sqrt{4m^2 - (2\dot{m}+1)}\,]x, \tag{46}$$

assuming, of course, that an optimal control exists.

CONCLUSION

It has been shown that the theory of Wonham [1] for minimum mean cost control can be extended to minimum cost variance problems in which the mean cost is predetermined. The main result is a Stochastic Hamilton-Jacobi equation which is necessary and sufficient for minimum variance controllers. A method for approaching the solution of this equation under the mean cost constraint is outlined by means of an example. The discussion is formal.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grant GK 804. The author wishes also to acknowledge several helpful conversations with Dr. R. Jeffrey Leake of the Department of Electrical Engineering, University of Notre Dame.

REFERENCES

1. W. M. Wonham, "Stochastic Problems in Optimal Control," 1963 IEEE International Convention Record, Part 2, pp. 114-124.

2. M. K. Sain, "On Minimal-Variance Control of Linear Systems with Quadratic Loss," Ph.D. Thesis, University of Illinois, 1965; also published as Technical Report R-240, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois; January, 1965.

3. M. K. Sain, "Control of Linear Systems According to the Minimal Variance Criterion - A New Approach to the Disturbance Problem," IEEE Transactions on Automatic Control, Vol. 11, No. 1, pp. 118-122; January, 1966; also published as Technical Report EE-656, University of Notre Dame, Notre Dame, Indiana; June 8, 1965.

4. M. K. Sain, "Relative Costs of Mean and Variance Control for a Class of Linear, Noisy Systems," Proceedings of the Third Annual Allerton Conference on Circuit and System Theory, Monticello, Illinois; 1965; pp. 121-129.

5. M. K. Sain and C. R. Souza, "A Projection Principle for Minimum Variance Estimation," Paper to be presented at 1966 National Electronics Conference, Chicago, Illinois.

OPTIMUM COMBINED CONTROL AND ESTIMATION FOR PARTIALLY CONTROLLED SYSTEMS

ROBERT E. LARSON
Stanford Research Institute
Menlo Park, California

ABSTRACT

In many optimal control problems the performance criterion depends on the behavior of a system that is not subject to control. Typical examples are tracking problems and intercept problems. It is known that these problems can be handled within the framework of optimal control theory by defining additional state variables to account for the dynamics of the uncontrolled systems and then appending these state variables to the state vector of the controlled system.

It is shown in this paper that the computation of optimum combined control and estimation for these problems can be broken into smaller calculations with a great saving in total computational requirements. Attention is focused on the linear case, where the known result of separation of control and estimation can be used directly. It is then shown that the control computation can be further separated into the calculation of optimal control for the uncontrolled system alone and then determination of a linear correction term. A separation within the estimation computation is also obtained. The extension of these results to certain nonlinear systems is indicated, and some practical problems to which these results have been applied are discussed.

A. INTRODUCTION

In many optimal control problems the performance criterion depends on the behavior of a system that is not subject to control. A typical example is the tracking of a satellite by a radar antenna; the satellite is not controlled, but the performance criterion depends on the error in angle between the antenna electrical axis and the coordinates of the satellite. Other examples are anti-missile missile control, in which the target to be intercepted cannot be controlled; production problems, in which the demand to be supplied is not controlled; and certain adaptive control problems.

A very useful viewpoint in attacking these problems is that the two systems actually form a single overall system, but only a part of this larger system can be controlled. In this section the optimum combined control and estimation[1*] of such systems is considered. Attention is focused on the case of linear system and measurement equations, white gaussian random variables, and a quadratic performance criterion. The well known result that the optimum estimation and optimal control functions are performed separately for this case can thus be used.[2,3] The major result obtained here is that the control function can be further separated. First, the optimal control signal for the controllable part of the system is computed using a Riccati equation. Next, a matrix linear equation is solved to obtain a term that accounts for the behavior of the uncontrolled part of the system. The optimal control for the overall system is then the sum of these two quantities. The major consequence of this result is that the problem of computing optimal control for an (m+n)-dimensional system, where m is the dimension of the uncontrolled system and n is the dimension of the controlled system, is reduced to computing optimal control for the n-dimen-

---

[*] References are listed at the end of this paper.

sional system. Since computational requirements for optimal control calculations increase exponentially with the dimension of the system, a great saving is obtained.

The estimation problem is also considered. It is shown that under certain conditions a similar separation is obtained.

Although these results are rigorously true only for the linear case described above, it is often possible to apply them to other problems by making a suitable linearization. A number of examples where such a linearization was made are described in Sec. G. In addition, it appears that similar separations can be made even in more general nonlinear problems.

B. PROBLEM FORMULATION

The controlled system is described by the linear difference equation

$$\underline{x}(k + 1) = \Phi_1(k)\ \underline{x}(k) + \Delta_1(k)\ \underline{u}(k) + \Gamma_1(k)\ \underline{w}_1(k), \tag{1}$$

where

$\underline{x}$ = State vector, n-dimensional
$\underline{u}$ = Control vector, q-dimensional
$\underline{w}_1$ = Random forcing vector, p-dimensional
Φ_1 = Transition matrix, n X n
Δ_1 = Distribution matrix, n X q
Γ_1 = Random forcing distribution matrix, n X p
k = Index of time.

The state is observed through the noisy linear measurement system.

$$\underline{z}_1(k) = H_1(k)\ \underline{x}(k) + \underline{v}_1(k) \tag{2}$$

where

$\underline{z}_1$ = Measurement vector, r-dimensional
$\underline{v}_1$ = Measurement noise vector, r-dimensional
H_1 = Observation matrix, r X n.

The uncontrolled system is also described by a linear difference equation

$$\underline{y}(k + 1) = \Phi_2(k)\ \underline{y}(k) + \Gamma_2(k)\ \underline{w}_2(k) \tag{3}$$

where

$\underline{y}$ = State vector, m-dimensional
$\underline{w}_2$ = Random forcing vector, p´-dimensional
Φ_2 = Transition matrix, m X m
Γ_2 = Random forcing distribution matrix, m X p´.

The control $\underline{u}$ has no effect on this system.

The state $\underline{y}$ is observed through a noisy linear measurement system,

$$\underline{z}_2(k) = H_2(k)\ \underline{y}(k) + \underline{v}_2(k) \tag{4}$$

where

$\underline{z}_2$ = Measurement vector, r´-dimensional
$\underline{v}_2$ = Measurement noise vector, r´-dimensional
H_2 = Observation matrix, r´ X m.

All the random variables are white and gaussian. They are completely characterized by their means and covariance matrices

$$E[\underline{w}_1(k)] = 0, \quad \mathrm{Cov}\ [\underline{w}_1(k)] = Q_1(k) \tag{5a}$$
$$E[\underline{w}_2(k)] = 0, \quad \mathrm{Cov}\ [\underline{w}_2(k)] = Q_2(k) \tag{5b}$$
$$E[\underline{v}_1(k)] = 0, \quad \mathrm{Cov}\ [\underline{v}_1(k)] = R_1(k) \tag{5c}$$

$$E]\underline{v}_2(k)] = 0, \quad \mathrm{Cov}\,[\underline{v}_2(k)] = R_2(k) \qquad k = 0, 1, \ldots, K. \tag{5d}$$

The initial states $\underline{x}(0)$ and $\underline{y}(0)$ are also gaussian random variables, characterized by their means and covariance matrices,

$$E[\underline{x}(0)] = \hat{\underline{x}}(0), \quad \mathrm{Cov}\,[\underline{x}(0)] = S_1(0) \tag{6a}$$

$$E[\underline{y}(0)] = \hat{\underline{y}}(0), \quad \mathrm{Cov}\,[\underline{y}(0)] = S_2(0). \tag{6b}$$

Finally, the performance criterion is the expected value of a quadratic time-varying function of all states:

$$J = E\left\{\sum_{k=0}^{K} [\underline{x}^T(k)\, A_1(k)\, \underline{x}(k) + 2\underline{x}^T(k)\, A_3(k)\, \underline{y}(k) + \underline{y}^T(k)\, A_2(k)\, \underline{y}(k) + \underline{u}^T(k)\, B(k)\, \underline{u}(k)]\right\} \tag{7}$$

where

A_1 = Quadratic cost matrix on states $\underline{x}$, n × n
A_3 = Cross-term cost matrix between states $\underline{x}$ and $\underline{y}$, n × m
A_2 = Quadratic cost matrix on states $\underline{y}$, m × m
B = Quadratic cost matrix on control, q × q

in which the expectation is over $\underline{x}(0)$, $\underline{y}(0)$, $\underline{w}_1(0), \ldots, \underline{w}_1(K)$, $\underline{w}_2(0), \ldots, \underline{w}_2(K)$, $\underline{v}_1(1), \ldots, \underline{v}_1(K)$, and $\underline{v}_2(1), \ldots, \underline{v}_2(K)$.

The problem can be stated as follows: Given (a) the system described by Eqs. (1) and (3); (b) the sequence of observations $\underline{z}_1(1), \ldots, \underline{z}_1(k)$, $\underline{z}_2(1), \ldots, \underline{z}_2(k)$; (c) the sequence of past controls $\underline{u}(0), \ldots, \underline{u}(k-1)$; (d) the covariance matrices $Q_1(0), \ldots, Q_1(k-1)$, $Q_2(0), \ldots, Q_2(k-1)$, $R_1(1), \ldots, R_1(k)$, $R_2(1), \ldots, R_2(k)$, $S_1(0)$, and $S_2(0)$, and the mean vectors $\hat{\underline{x}}(0)$ and $\hat{\underline{y}}(0)$.

Find the sequence of controls $\underline{u}(k), \ldots, \underline{u}(K)$ that minimizes

$$J = E\left\{\sum_{j=k}^{K} [\underline{x}^T(j)\, A_1(j)\, \underline{x}(j) + \underline{x}^T(j)\, A_3(j)\, \underline{y}(j) + \underline{y}^T(j)\, A_2(j)\, \underline{y}(j) + \underline{u}^T(j)\, B(j)\, \underline{u}(j)]\right\} \tag{8}$$

where the expectation is taken over $\underline{w}_1(k), \ldots, \underline{w}_1(K)$, $\underline{w}_2(k), \ldots, \underline{w}_2(K)$, $\underline{v}_1(k+1), \ldots, \underline{v}_1(K)$, $\underline{v}_2(k+1), \ldots, \underline{v}_2(K)$, and $\underline{x}(k)$, conditioned on $\underline{z}_1(1), \ldots, \underline{z}_1(k)$, $\underline{z}_2(1), \ldots, \underline{z}_2(k)$, and $\underline{u}(0), \ldots, \underline{u}(k-1)$.

The problem can be put into a slightly more compact form by writing a composite state vector,

$$\begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix}.$$

The system equations are then

$$\begin{bmatrix} \underline{x}(k+1) \\ \underline{y}(k+1) \end{bmatrix} = \begin{bmatrix} \Phi_1(k) & 0 \\ 0 & \Phi_2(k) \end{bmatrix} \begin{bmatrix} \underline{x}(k) \\ \underline{y}(k) \end{bmatrix} + \begin{bmatrix} \Delta_1(k) \\ 0 \end{bmatrix} \underline{u}(k) + \begin{bmatrix} \Gamma_1(k) & 0 \\ 0 & \Gamma_2(k) \end{bmatrix} \begin{bmatrix} \underline{w}_1(k) \\ \underline{w}_2(k) \end{bmatrix} . \tag{9}$$

The observation equations are

$$\begin{bmatrix} \underline{z}_1(k) \\ \underline{z}_2(k) \end{bmatrix} = \begin{bmatrix} H_1(k) & 0 \\ 0 & H_2(k) \end{bmatrix} \begin{bmatrix} \underline{x}(k) \\ \underline{y}(k) \end{bmatrix} + \begin{bmatrix} \underline{v}_1(k) \\ \underline{v}_2(k) \end{bmatrix} . \tag{10}$$

The random variables are characterized by

$$\left\{ \begin{bmatrix} \underline{w}_1(k) \\ \underline{w}_2(k) \end{bmatrix} \right\} = 0 \quad , \quad \text{Cov} \left\{ \begin{bmatrix} \underline{w}_1(k) \\ \underline{w}_2(k) \end{bmatrix} \right\} = \begin{bmatrix} Q_1(k) & 0 \\ 0 & Q_2(k) \end{bmatrix} \tag{11a}$$

$$E \left\{ \begin{bmatrix} \underline{v}_1(k) \\ \underline{v}_2(k) \end{bmatrix} \right\} = 0 \quad , \quad \text{Cov} \left\{ \begin{bmatrix} \underline{v}_1(k) \\ \underline{v}_2(k) \end{bmatrix} \right\} = \begin{bmatrix} R_1(k) & 0 \\ 0 & R_2(k) \end{bmatrix} \tag{11b}$$

$$E \left\{ \begin{bmatrix} \underline{x}(0) \\ \underline{y}(0) \end{bmatrix} \right\} = \begin{bmatrix} \hat{\underline{x}}(0) \\ \hat{\underline{y}}(0) \end{bmatrix} \quad , \quad \text{Cov} \left\{ \begin{bmatrix} \underline{x}(0) \\ \underline{y}(0) \end{bmatrix} \right\} = \begin{bmatrix} S_1(0) & 0 \\ 0 & S_2(0) \end{bmatrix} . \tag{11c}$$

The performance criterion is

$$J = E \left\{ \sum_{k=0}^{K} [\underline{x}^T(k) \ \underline{y}^T(k)] \begin{bmatrix} A_1(k) & A_3(k) \\ A_3^T(k) & A_2(k) \end{bmatrix} \begin{bmatrix} \underline{x}(k) \\ \underline{y}(k) \end{bmatrix} + \underline{u}^T(k) \ B(k) \ \underline{u}(k) \right\} . \tag{12}$$

where the expectation is as in Eq. (8).

C. SEPARATION OF CONTROL AND ESTIMATION

It has been shown[2,3] that if the system and measurement equations are linear; if all random variables are gaussian, white and zero-mean; and if the performance criterion is quadratic, then the solution to the combined optimal control and estimation is obtained by first generating an optimum estimate of the present state and then solving the optimum control problem as if this estimate were in fact the actual state. The optimum estimate is obtained from the Kalman filter equations,[4] while the optimal control is found by solving a Riccati equation.[5]

For the system equations,

$$\underline{x}^*(k+1) = \Phi^*(k) \ \underline{x}^*(k) + \Delta^*(k) \ \underline{u}^*(k) + \Gamma^*(k) \ \underline{w}^*(k) \tag{13}$$

the measurement equations,

$$\underline{z}^*(k) = H^*(k) \ \underline{x}^*(k) + \underline{v}^*(k) \tag{14}$$

and the white, gaussian random variables characterized by

$$E[\underline{w}^*(k)] = 0 \quad , \quad \text{Cov} \ [\underline{w}^*(k)] = Q^*(k) \tag{15a}$$

$$E[\underline{v}^*(k)] = 0 \quad , \quad \text{Cov} \ [\underline{v}^*(k)] = R^*(k) \tag{15b}$$

$$E[\underline{x}^*(0)] = \hat{\underline{x}}^*(0) \quad , \quad \text{Cov} \ [\underline{x}^*(0)] = S^*(0) \tag{15c}$$

it can be shown that the Kalman estimate, $\hat{\underline{x}}^*(k|k)$, obeys the recursive equation,

$$\hat{\underline{x}}^*(k + 1|k + 1) = \Phi^*(k)\ \hat{\underline{x}}^*(k|k) + \Delta^*(k)\ \underline{u}^*(k)$$
$$+ W^*(k + 1)[\underline{z}^*(k + 1) - H^*(k + 1)\ \Phi^*(k)\ \hat{\underline{x}}^*(k|k)] \quad (16)$$

where

$$W^*(k + 1) = S^*(k + 1|k)\ H^{*T}(k + 1)$$
$$\times\ [R^*(k + 1) + H^*(k + 1)\ S^*(k + 1|k)\ H^{*T}(k + 1)]^{-1}$$

$$S^*(k + 1|k) = \Phi^*(k)\ S^*(k|k)\ \Phi^{*T}(k) + \Gamma^*(k)\ Q^*(k)\ \Gamma^{*T}(k) \quad .$$

For use at the next time instant, $S^*(k + 1|k + 1)$ is computed as

$$S^*(k + 1|k + 1) = [I - W^*(k)\ H^*(k)]\ S^*(k + 1|k) \quad . \quad (17)$$

As initial conditions for these equations,

$$\hat{\underline{x}}^*(0|0) = \hat{\underline{x}}^*(0) \quad (18a)$$

$$S^*(0|0) = S^*(0) \quad (18b)$$

For the system equations in Eq. (13) and the performance criterion

$$J = \sum_{k=0}^{K} [\underline{x}^{*T}(k)\ A^*(k)\ \underline{x}^*(k) + \underline{u}^{*T}(k)\ B^*(k)\ \underline{u}^*(k)] \quad (19)$$

the optimal control, $\hat{\underline{u}}^*(\underline{x}^*, k)$, takes the form,

$$\hat{\underline{u}}^*(\underline{x}^*, k) = - G^*(k)\ \underline{x}^* \quad . \quad (20)$$

The gain matrix G^* is given by

$$G^*(k) = [B^*(k) + \Delta^{*T}(k)\ P^*(k + 1)\ \Delta^*(k)]^{-1}\ \Delta^*(k)\ P^*(k + 1)\ \Phi^*(k) \quad . \quad (21)$$

The matrix $P^*(k)$ obeys the discrete Riccati equation,

$$P^*(k) = A^*(k) + \Phi^{*T}(k)\ P^*(k + 1)\ [\Phi^*(k) - \Delta^*(k)\ G^*(k)]$$
$$= A^*(k) + \Phi^{*T}(k)\ P^*(k + 1)\ \Phi^*(k) - \Phi^{*T}(k)\ P^*(k + 1)\ \Delta^*(k)$$
$$\times\ [B^*(k) + \Delta^{*T}(k)\ P^*(k + 1)\ \Delta^*(k)]^{-1}\ \Delta^T(k)\ P^*(k + 1)\ \Phi^*(k) \quad . \quad (22)$$

Note that this equation runs backward in time. Customarily, $B^*(K) = 0$, so that as a starting condition

$$P^*(K) = A^*(K) \quad . \quad (23)$$

At a given value of k, the solution to the combined optimization problem is found by calculating the estimate $\hat{\underline{x}}^*(k|k)$ from iterative use of Eq. (16) and the feedback gain $G^*(k)$ from Eq. (21) and iterative use of Eq. (22). The optimal control is then given by Eq. (20) as

$$\hat{\underline{u}}^* = - G^*(k)\ \hat{\underline{x}}^*(k|k) \quad . \quad (24)$$

Since there is no overlap in the calculation of $G^*(k)$ and $\hat{\underline{x}}^*(k|k)$, the problems of estimation and control can indeed be solved separately.

D. SEPARATION WITHIN THE CONTROL COMPUTATION

The computation of optimal control for the overall system described by Eqs. (9) through (12) will now be considered.

Making the identifications

$$\underline{x}^*(k) = \begin{bmatrix} \underline{x}(k) \\ \underline{y}(k) \end{bmatrix} \quad (25a)$$

$$\Phi^*(k) = \begin{bmatrix} \Phi_1(k) & 0 \\ 0 & \Phi_2(k) \end{bmatrix} \tag{25b}$$

$$\Delta^*(k) = \begin{bmatrix} \Delta_1(k) \\ 0 \end{bmatrix} \tag{25c}$$

$$\Gamma^*(k) = \begin{bmatrix} \Gamma_1(k) & 0 \\ 0 & \Gamma_2(k) \end{bmatrix} \tag{25d}$$

$$A^*(k) = \begin{bmatrix} A_1(k) & A_3(k) \\ A_3^T(k) & A_2(k) \end{bmatrix} \tag{25e}$$

$$B^*(k) = B(k) \tag{25f}$$

$$\underline{u}^*(k) = \underline{u}(k) \tag{25g}$$

$$\underline{w}^*(k) = \begin{bmatrix} \underline{w}_1(k) \\ \underline{w}_2(k) \end{bmatrix} \tag{25h}$$

$$Q^*(k) = \begin{bmatrix} Q_1(k) & 0 \\ 0 & Q_2(k) \end{bmatrix} \tag{25i}$$

Eqs. (21) and (22) can be used directly. It is convenient in this problem to partition $P^*(k)$ so that it has the same form as $A^*(k)$:

$$P^*(k) = \begin{bmatrix} P_1(k) & P_3(k) \\ P_3^T(k) & P_2(k) \end{bmatrix} \quad . \tag{26}$$

The matrix whose inverse appears in Eq. (21) can be calculated as

$$\begin{aligned} &[B^*(k) + \Delta^{*T}(k)\ P^*(k+1)\ \Delta^*(k)] \\ &=[B(k)] + [\Delta_1^T(k)\ \ 0] \begin{bmatrix} P_1(k+1) & P_3(k+1) \\ P_3^T(k+1) & P_2(k+1) \end{bmatrix} \begin{bmatrix} \Delta_1(k) \\ 0 \end{bmatrix} \\ &=[B(k) + \Delta_1^T(k)\ P_1(k+1)\ \Delta_1(k)] \quad . \end{aligned} \tag{27}$$

The other matrix in Eq. (21) is obtained as

$$\begin{aligned} \Delta^{*T}(k)\ P^*(k+1)\ \Phi^*(k) &= [\Delta_1^T(k)\ \ 0] \begin{bmatrix} P_1(k+1) & P_3(k+1) \\ P_3^T(k+1) & P_2(k+1) \end{bmatrix} \begin{bmatrix} \Phi_1(k) & 0 \\ 0 & \Phi_2(k) \end{bmatrix} \\ &= [\Delta_1^T(k)\ P_1(k+1)\ \Phi_1(k) \quad \Delta_1^T(k)\ P_3(k+1)\ \Phi_2(k)] \quad . \end{aligned} \tag{28}$$

The optimal control gain matrix is then

$$\begin{aligned} G^*(k) &= [B^*(k) + \Delta^{*T}(k)\ P^*(k+1)\ \Delta^*(k)]^{-1}\ \Delta^{*T}(k)\ P^*(k+1)\ \Phi^*(k) \\ &= [B(k) + \Delta_1^T(k)\ P_1(k+1)\ \Delta_1(k)]^{-1} \\ &\quad \times [\Delta_1^T(k)\ P_1(k+1)\ \Phi_1(k)\ \ .\ \ \Delta_1^T(k)\ P_3(k+1)\ \Phi_2(k)] \quad . \end{aligned} \tag{29}$$

The optimal control is thus

$$\begin{aligned} \hat{\underline{u}} = &- [B(k) + \Delta_1^T(k)\ P_1(k+1)\ \Delta_1(k)]^{-1}\ \Delta_1^T(k)\ P_1(k+1)\ \Phi_1(k)\ \hat{\underline{x}}(k|k) \\ &- [B(k) + \Delta_1^T(k)\ P_1(k+1)\ \Delta_1(k)]^{-1}\ \Delta_1^T(k)\ P_3(k+1)\ \Phi_2(k)\ \hat{\underline{y}}(k|k) \quad . \end{aligned} \tag{30}$$

The first term in the control equation is exactly the control that would be obtained if the system equation involved $\underline{x}$ and $\underline{u}$ only [i.e., if the system equation were Eq. (1) and if the performance criterion were a quadratic function only of $\underline{x}$ and $\underline{u}$]. The second term involves $\hat{\underline{y}}(k|k)$, and hence can be regarded as a correction term to account for the coupling introduced through the performance criterion. Eq. (7). Note that this term involves $P_1(k + 1)$ and $P_3(k + 1)$, but not $P_2(k + 1)$.

The recursive equation for $P^*(k)$ can also be found by substitution. It is straightforward to obtain that

$$\Phi^{*T}(k)P^*(k + 1)\Phi^*(k) = \begin{bmatrix} \Phi_1^T(k)P_1(k + 1)\Phi_1(k) & \Phi_1^T(k)P_3(k + 1)\Phi_2(k) \\ \Phi_2^T(k)P_3^T(k + 1)\Phi_1(k) & \Phi_2^T(k)P_2(k + 1)\Phi_2(k) \end{bmatrix} \tag{31}$$

and

$$\Phi^{*T}(k)\ P^*(k + 1)\ \Delta^*(k) = \begin{bmatrix} \Phi_1^T(k)\ P_1(k + 1)\ \Delta_1(k) \\ \Phi_1^T(k)\ P_3^T(k + 1)\ \Delta_1(k) \end{bmatrix} \quad . \tag{32}$$

Combining Eqs. (27), (28), (31), and (32) and substituting into Eq. (22), the complete recursive equation for $P^*(k)$ can be obtained. The resulting matrix equation can be partitioned into separate equations for $P_1(k)$, $P_2(k)$, and $P_3(k)$. The equation for $P_3^T(k)$ is, of course, simply the transpose of the equation for $P_3(k)$. The results are

$$\begin{aligned} P_1(k) &= A_1(k) + \Phi_1^T(k)\ P_1(k + 1)\ \Phi_1(k) - \Phi_1^T(k)\ P_1(k + 1)\ \Delta_1(k) \\ &\quad \times [B(k) + \Delta_1^T(k)\ P_1(k + 1)\ \Delta_1(k)]^{-1}\ \Delta_1^T(k)\ P_1(k + 1)\ \Phi_1(k) \quad , \end{aligned} \tag{33}$$

$$\begin{aligned} P_3(k) &= A_3(k) + \Phi_1^T(k)\ P_3(k + 1)\ \Phi_2(k) - \Phi_1^T(k)\ P_1(k + 1)\ \Delta_1(k) \\ &\quad \times [B(k) + \Delta_1^T(k)\ P_1(k + 1)\ \Delta_1(k)]^{-1}\ \Delta_1^T(k)\ P_3(k + 1)\ \Phi_2(k) \quad , \end{aligned} \tag{34}$$

$$\begin{aligned} P_2(k) &= A_2(k) + \Phi_2^T(k)\ P_2(k + 1)\ \Phi_2(k) - \Phi_2^T(k)\ P_3^T(k + 1)\ \Delta_1(k) \\ &\quad \times [B(k) + \Delta_1^T(k)\ P_1(k + 1)\ \Delta_1(k)]^{-1}\ \Delta_1^T(k)\ P_3(k + 1)\ \Phi_2(k) \quad . \end{aligned} \tag{35}$$

It can be seen immediately that Eq. (33) is the Riccati equation for the problem with a system equation involving $\underline{x}$ and $\underline{u}$ only--i.e., the system equation is Eq. (5)--and with performance criterion

$$J = \sum_{k=0}^{K} [\underline{x}^T(k)\ A_1(k)\ \underline{x}(k) + \underline{u}^T(k)\ B(k)\ \underline{u}(k)] \quad .$$

Equation (33) depends only on Φ_1, Δ_1, A_1, and B, hence it can be solved independently of the other components of P. Thus, the dimension of the Riccati equation that must be solved has been reduced from (m + n) × (m + n) to n × n, the dimension of $P_1(k)$.

It can also be seen that once $P_1(k)$ has been found, Eq. (34) is a linear equation in $P_3(k)$, not a Riccati equation. The dimension of $P_3(k)$ is m × n, again much less than (m + n) × (m + n).

Finally, since $P_2(k)$ has no effect on either the control or the recursive equations for $P_3(k)$ and $P_1(k)$, it need not be computed at all.

E. COMPUTATIONAL ASPECTS

The computation of optimal control is most efficiently accomplished as follows. First, the matrix $C_1(k)$, defined as

$$C_1(k) = [B(k) + \Delta_1^T(k)\ P_1(k + 1)\ \Delta_1(k)]^{-1}\ \Delta_1^T(k) \tag{36}$$

is computed. The optimal control, $\hat{\underline{u}}$, is then found as

$$\hat{\underline{u}} = - C_1(k)\ [P_1(k + 1)\ \Phi_1(k)\ \hat{\underline{x}}(k|k) + P_3(k + 1)\ \Phi_2(k)\ \hat{\underline{y}}(k|k)] \quad . \tag{37}$$

The matrix $C_2(k)$, where

$$C_2(k) = \Phi_1(k) \, [I - P_1(k + 1) \, \Delta_1(k) \, C_1(k)] \tag{38}$$

is next computed. Then, Eqs. (33) and (34) can be rewritten as

$$P_1(k) = A_1(k) + C_2(k) \, P_1(k + 1) \, \Phi_1(k) \tag{39}$$

$$P_3(k) = A_3(k) + C_2(k) \, P_3(k + 1) \, \Phi_2(k) \quad . \tag{40}$$

Since the matrix products $P_1(k + 1) \, \Phi_1(k)$ and $P_3(k + 1) \, \Phi_2(k)$ were computed in Eq. (37), the calculation of these terms in Eqs. (39) and (40) need not be repeated.

In many problems it is desired to compute the steady-state optimal control--i.e., the optimal control that would be applied if the matrices Φ^*, Γ^*, Δ^*, A^*, and B^* are all time-invariant and if K, the final time, is allowed to approach infinity. In this case it can be shown that P^* approaches a constant matrix,[5] which can be found by setting

$$P^* = P^*(k) = P^*(k + 1) \tag{41}$$

in Eq. (22). The steady-state optimal control is then obtained by substituting this value of P^* into Eqs. (20) and (21).

Because Eq. (22) is a Riccati equation, the equations for the components of P^* are nonlinear. If P^* is of high dimension, the most effective method of solution is to make an initial guess for P^*, substitute this value wherever $P^*(k + 1)$ appears on the right-hand side of Eq. (22), solve for $P^*(k)$, and repeat until $P^*(k)$ approaches $P^*(k + 1)$. This is often a difficult and time-consuming task.

If Eqs. (33) and (34) are used to find the steady-state optimal control rather than Eqs. (20) through (22), a great saving in computational requirements is again obtained. Since Eq. (33) is a Riccati equation in $P_1(k)$, it can be solved for the steady-state value by the same method as above. However, the dimension of P_1 is $n \times n$ rather than $(m + n) \times (m + n)$. Once P_1 has been found, it can be substituted into Eq. (34). This latter equation is now a linear equation in P_3, not a Riccati equation. These matrices can then be substituted into Eq. (30) to obtain the optimal steady-state control. The result is that the solution of an $(m + n) \times (m + n)$ steady-state Riccati equation has been reduced to the solution of an $n \times n$ steady-state Riccati equation followed by the solution of an $n \times m$ linear matrix equation.

If the matrix Φ_2 is the identity matrix, then the solution to the linear equation for P_3 is particularly simple. The terms involving P_3 can be collected on the left side of the equation, with the result that A_3 is equal to P_3 pre-multiplied by some matrix. The inversion of this matrix yields

$$P_3 = [I - \Phi_1^T + \Phi_1^T \, P_1 \, \Delta_1 \, (B + \Delta_1^T \, P_1 \, \Delta_1)^{-1} \, \Delta_1^T]^{-1} \, A_3 \quad . \tag{42}$$

F. SEPARATION WITHIN THE ESTIMATION COMPUTATION

If the identifications

$$\underline{x}^*(k) = \begin{bmatrix} \underline{x}(k) \\ \underline{y}(k) \end{bmatrix} \tag{43a}$$

$$\Phi^*(k) = \begin{bmatrix} \Phi_1(k) & 0 \\ 0 & \Phi_2(k) \end{bmatrix} \tag{43b}$$

$$\Delta^*(k) = \begin{bmatrix} \Delta_1(k) \\ 0 \end{bmatrix} \tag{43c}$$

$$\underline{w}^*(k) = \begin{bmatrix} \underline{w}_1(k) \\ \underline{w}_2(k) \end{bmatrix} \tag{43d}$$

$$Q^*(k) = \begin{bmatrix} Q_1(k) & 0 \\ 0 & Q_2(k) \end{bmatrix} \tag{43e}$$

$$\Gamma^*(k) = \begin{bmatrix} \Gamma_1(k) & 0 \\ 0 & \Gamma_2(k) \end{bmatrix} \tag{43f}$$

$$\underline{z}^*(k) = \begin{bmatrix} \underline{z}_1(k) \\ \underline{z}_2(k) \end{bmatrix} \tag{43g}$$

$$H^*(k) = \begin{bmatrix} H_1(k) & 0 \\ 0 & H_2(k) \end{bmatrix} \tag{43h}$$

$$\underline{v}^*(k) = \begin{bmatrix} \underline{v}_1(k) \\ \underline{v}_2(k) \end{bmatrix} \tag{43i}$$

$$R^*(k) = \begin{bmatrix} R_1(k) & 0 \\ 0 & R_2(k) \end{bmatrix} \tag{43j}$$

$$\hat{\underline{x}}^*(0) = \begin{bmatrix} \hat{\underline{x}}(0) \\ \hat{\underline{y}}(0) \end{bmatrix} \tag{43k}$$

$$S^*(0) = \begin{bmatrix} S_1(0) & 0 \\ 0 & S_2(0) \end{bmatrix} \tag{43l}$$

are made, then it is straightforward to show, using Eqs. (13) through (18), that the estimation procedure separates into

$$\begin{aligned} \hat{\underline{x}}(k+1|k+1) &= \Phi_1(k)\,\hat{\underline{x}}(k|k) + \Delta_1(k)\,\underline{u}(k) + W_1(k+1) \\ &\quad \times [\underline{z}_1(k+1) - H_1(k+1)\,\Phi_1(k)\,\hat{\underline{x}}(k|k)] \end{aligned} \tag{44}$$

where

$$\begin{aligned} W_1(k+1) &= S_1(k+1|k)\,H_1^T(k+1) \\ &\quad \times [R_1(k+1) + H_1(k+1)\,S_1(k+1|k)\,H_1^T(k+1)]^{-1} \\ S_1(k+1|k) &= \Phi_1(k)\,S_1(k|k)\,\Phi_1^T(k) + \Gamma_1(k)\,Q_1(k)\,\Gamma_1^T(k) \end{aligned}$$

and

$$\begin{aligned} \hat{\underline{y}}(k+1|k+1) &= \Phi_2(k)\,\hat{\underline{y}}(k|k) + W_2(k+1) \\ &\quad \times [\underline{z}_2(k+1) - H_2(k+1)\,\Phi_2(k)\,\hat{\underline{y}}(k|k)] \end{aligned} \tag{45}$$

where

$$\begin{aligned} W_2(k+1) &= S_2(k+1|k)\,H_2^T(k+1) \\ &\quad \times [R_2(k+1) + H_2(k+1)\,S_2(k+1|k)\,H_2^T(k+1)]^{-1} \\ S_2(k+1|k) &= \Phi_2(k)\,S_2(k|k)\,\Phi_1^T(k) + \Gamma_2(k)\,Q_2(k)\,\Gamma_2^T(k) \quad . \end{aligned}$$

For use at the next time instant, $S_1(k+1|k+1)$ and $S_2(k+1|k+1)$ are computed as

$$S_1(k + 1|k + 1) = [I - W_1(k)\ H_1(k)]\ S_1(k + 1|k) \tag{46}$$

$$S_2(k + 1|k + 1) = [I - W_2(k)\ H_2(k)]\ S_2(k + 1|k)\quad . \tag{47}$$

As initial conditions,

$$\hat{\underline{x}}(0|0) = \hat{\underline{x}}(0)\quad ,\quad S_1(0|0) = S_1(0) \tag{48}$$

$$\hat{\underline{y}}(0|0) = \hat{\underline{y}}(0)\quad ,\quad S_2(0|0) = S_2(0) \tag{49}$$

G. APPLICATIONS

One class of problem to which these results apply is the servo or tracking problem. In this case the objective is to make a given system follow or track the state of a second system. It is well known that the servo problem can be converted to a regulator problem by defining additional state variables to account for the dynamics of the system being tracked.[2] What the results of Sec D signify is that the dimension of the problem to be solved is not the sum of the dimensions of the two systems, but instead is on the order of the dimension of the system that is controlled.

The servo problem is generally formulated as follows. If $\underline{x}$ is the state vector for the controlled system and $\underline{y}$ the state vector for the system to be tracked, and if the objective is to track a vector $L_2\underline{y}$ by the vector $L_1\underline{x}$, then the performance criterion can be written as a quadratic function of the error between these two vectors:

$$J = \sum_{k=0}^{K} \{[L_1(k)\ \underline{x}(k) - L_2(k)\ \underline{y}(k)]^T\ M(k)[L_1(k)\ \underline{x}(k) - L_2(k)\ \underline{y}(k)] + \underline{u}^T(k)\ B(k)\ \underline{u}(k)\}\quad . \tag{50}$$

This criterion then takes the form of Eq. (12) with

$$\begin{bmatrix} A_1(k) & A_3(k) \\ A_3^T(k) & A_2(k) \end{bmatrix} = \begin{bmatrix} L_1^T(k)\ M(k)\ L_1(k) & -\ L_1^T(k)\ M(k)\ L_2(k) \\ -\ L_2^T(k)\ M(k)\ L_1(k) & L_2^T(k)\ M(k)\ L_2(k) \end{bmatrix}\quad . \tag{51}$$

If the remaining equations for $\underline{x}$ and $\underline{y}$ are linear, then the results of Sec. D can be used directly.

In the classical approach[6] to this problem the state variables are chosen to be the components of the error vectors $[L_1(k)\ \underline{x}(k) - L_2(k)\ \underline{y}(k)]$, and the optimal control is assumed to be a linear function of these states. It can be seen from Eq. (30) that this is true if, and only if, there exists a matrix $C(k)$ such that

$$C(k)\ [L_1(k)\underline{x}(k) - L_2(k)\underline{y}(k)] = \Delta_1^T(k)\ [P_1(k + 1)\Phi_1(k)\underline{x}(k) + P_3(k + 1)\Phi_2(k)\underline{y}(k)]. \tag{52}$$

If $\underline{x}$ and $\underline{y}$ have the same dimension, if $\Phi_1(k) = \Phi_2(k)$, and if $L_1(k) = L_2(k)$, it can be shown that $P_1(k) = -\ P_3(k)$, and the result in Eq. (52) then follows. Other cases under which Eq. (52) is satisfied can be found, but it is clear that in general the optimal control is not a linear function of this error.

A specific application ot which these results have been applied is an antenna tracking problem.[7] The equations for the antenna were linear, while the satellite equations were linearized. In this case the measurement equations are such that the estimation function cannot be separated as in Sec. F, but the control equations are such that the results of Sec. D apply.

Another class of problems to which the results apply is pursuit or intercept problems. The results have been used in a study of the NIKE-X system.[8] The state vector for the incoming target is $\underline{y}(k)$, the state vector for the anti-missile missile (AMM) is $\underline{x}(k)$, and the control vector for the AMM is $\underline{u}(k)$. In this case the equations for both the target and the AMM must be linearized. The performance criterion is the expected value of a quadratic function of miss distance. In this case the measurement equations are such that estimation can be separated as in Sec. F, and control can be separated as in Sec. D.

A third class of problem involves the computation of optimal control decisions so that a time-varying demand is met. Many industrial and operation research problems, such as inventory control, airline scheduling, short-term gas pipeline optimization and electric power dispatching, are of this type. The results of Secs. D and F can be applied by defining $\underline{y}(k)$ to be the state vector for the dynamic process generating the demand, $\underline{x}(k)$ the state vector for the system that is controlled, and $\underline{u}(k)$ the vector of control decisions. Specific applications are currently under study.

Finally, the application of this result to a class of adaptive control problems is indicated in Ref. 7.

REFERENCES

1. L. Meier, "Combined Optimum Control and Estimation Theory," Contractor Report, SRI Project 5237, Contract NAS 2-2457, Stanford Research Institute, Menlo Park, California (October 1965).

2. T. L. Gunckel, II, "Optimum Design of Sampled-Data Systems with Random Parameters," TR 2102-2, Contracts Nonr-225(24) and NR 373 360, Stanford Electronics Laboratories, Stanford, California (24 April 1961).

3. P. D. Joseph and J. T. Tou, "On Linear Control Theory," Trans. AIEE, Part II 80, p. 18, No. 56, pp. 193-196 (September 1961).

4. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, J. Basic Engineering, pp. 35-45 (March 1960).

5. R. E. Kalman and R. W. Koepcke, "Optimal Synthesis of Linear Sampling Control Systems Using Generalized Performance Indexes," Trans. ASME, Vol. 80, No. 6, pp. 1820-1838 (November 1958).

6. G. C. Newton, L. A. Gould, and J. F. Kaiser, Analytical Design of Linear Feedback Controls, John Wiley and Sons, Inc., New York, 1957.

7. R. E. Larson, R. M. Dressler, and J. Peschon, "Research on the Design of Adaptive Control Systems," Quarterly Report 3, SRI Project 5578, NASA Contract NAS 12-59, Stanford Research Institute, Menlo Park, California (April, 1966).

8. J. Peschon and R. E. Larson, "Analysis of an Intercept System," Final Report and Appendices, Contract DA-01-021-AMC-90006(Y), SRI Project 5188-7, Stanford Research Institute, Menlo Park, California (December 1965).

LOCALLY OPTIMUM DIFFERENTIAL GAMES

CELSO DE RENNA E SOUZA
MICHAEL D. CILETTI
University of Notre Dame
Notre Dame, Indiana

ABSTRACT

A direct application of the theory of differential games to practical problems is hampered by the requirement that globally optimum strategies be obtained by retrogressive integration of a characteristic set of differential equations, the boundary conditions of which are specified on a target surface in the player space. In many cases of practical importance, a global target surface with appropriate boundary conditions on the state and adjoint variables is difficult to define. Furthermore, global solutions present formidable storage problems for computer facilities available to the players of the game.

In this paper, a continuous succession of "locally optimum games" is used to determine a global sub-optimal strategy to be used in lieu of a solution to the globally optimum game. The locally optimum games are, trivially, asymptotically globally optimum; a substantial simplification of the boundary conditions is shown to result from the new formulation of the problem. Pertinent examples illustrate the application of the locally optimum criterion to problems of differential games and of their weighty particular case, optimal controls.

Motivation

Conventional approaches for multistage decision processes have a common point: the solution is constructed by induction or integration from the game terminus, backwards. The principle of optimality (see for instance R. Aris [1]) dominates the conceptual approaches. While invaluable for the a priori design of control systems, it is inefficient and even impossible to apply to certain games which are played as Von Neumann first visualized them, [2] strictly forward. Even in the continuous case, the remoteness of and the difficulty of defining a terminal surface makes the approach cumbersome when real-time computations are required from a vehicle-carried computer.

The mechanism of human thought when faced with similar problems has been thoroughly studied particularly in the area of chess-playing machines. The methods and approach of the great masters have been empirically studied and compiled [3], the thought process analyzed [4], and computer programs written [5]. The search seems to be done forward, down the choice tree, and the choice among the possible moves is based on various factors. Historically celebrated mating strategies were found by considering trees with about 100 positions, while, in general, even in complicated situations, the masters rarely generate a tree of more than 50 positions. The heuristic used to guide the players' choice is rather complicated and varies from player to player, even changing during a certain partie. However, all of them must have a common feature; they aid at putting the player at the most favorable position for subsequent moves, and to restrict the number of possible replies by the opponent.

Such reasoning seems to adapt itself readily to both discrete and continuous game models, and is the mainstay of the development to follow.

The Discrete Model

For a better understanding of the continuous model and for application in typically discrete cases, the description of a time-discrete game of prescribed duration will be used, after Fleming [6].

Let $j\Delta n = T$.

The game will have j moves with $i=1,2,\ldots j$ labelling them.

At the ith move player I chooses $y_i \varepsilon Y \subset R^{\lambda}$: II choses $Z_i \varepsilon Z \subset R^{K}$, etc., and the game is of perfect information, with simultaneous moves. The new state is given by:

$$x_{i+1} = x_i + \Delta n \cdot g(x_i, y_i, z_i) \tag{1}$$

The total payoff for I is

$$F_j(x) = \sum_{i=1}^{j} \Delta n f(x_i, y_i, z_i) + V^{\circ}(x_{j+1}) \tag{2}$$

with $V^{\circ}(x_{j+1})$ as terminal pay-off and satisfying a uniform Lipchitz condition in x on a bounded subset of the state space, ξ, and with constant K'. The value function is defined by induction

$$\begin{aligned} V(x_1, j\Delta n) &= \underset{y,z}{\text{val}}\; F_j(x_1) = \sum_{i=1}^{j} nf(x_i^*; y_i^*; z_i^*) + V^{\circ}(x_{j+1}^*) \\ &= \Delta n f(x_1, y_1^*, z_1^*) + V[x_1 + g(x_1, y_1^*, z_1^*); (j-1)\Delta n] \end{aligned} \tag{3}$$

where $\underset{y,z}{\text{val}}$ stands for $\underset{z}{\text{Min}}\;\underset{y}{\text{Max}}$ in the zero-sum game, $\underset{z}{\text{Max}}\;\underset{y}{\text{Max}}$ in the fully co-operative game, and where the optimizing control tactics are

$$\begin{bmatrix} y_1^* \\ \cdot \\ \cdot \\ \cdot \\ y_j^* \end{bmatrix} \text{ and } \begin{bmatrix} z_1^* \\ \cdot \\ \cdot \\ \cdot \\ z_j^* \end{bmatrix}, \text{ and the states reached: } \begin{bmatrix} x_1^* \\ \cdot \\ \cdot \\ \cdot \\ x_{j+1}^* \end{bmatrix}$$

Now consider a sequence of $(j-k)$ games of duration $k\Delta n$. For each of these the pay-off will be:

$$F_k^{\ell}(x) = \sum_{i=1}^{k} \Delta n f(x_i, y_i, z_i) + \mathring{P}(x_{k+1}^{\ell}) \quad \ell = 1, \ldots, (j-k) \tag{4}$$

and such that x_1 is the starting point for $\ell=1$ with $\mathring{P}(x)$ a preference function, defined on ξ and coinciding with $\mathring{V}(x_{j+1})$ on the terminal positions on the jth game but not necessarily so for intermediate states.

Let the local tactics be built in the following manner: use the first optimal decisions for k-games with $\ell=1,\ldots,(j-k-1)$ and all k optimal decisions of the k-game with $\ell=(j-k)$. We shall denote this by:

$$\begin{bmatrix} y_1' \\ \cdot \\ \cdot \\ \cdot \\ y_j' \end{bmatrix} \text{ and } \begin{bmatrix} z_1' \\ \cdot \\ \cdot \\ \cdot \\ z_j' \end{bmatrix}$$

and the states reached in this play:

$$\begin{bmatrix} x'_1 \\ x'_2 \\ \cdot \\ \cdot \\ \cdot \\ x'_{j+1} \end{bmatrix}$$

with:

$$x'_1 = x^*_1 \tag{5}$$

and: $$\overset{\circ}{P}(x'_{j+1}) = \overset{\circ}{V}(x'_{j+1}) \tag{6}$$

An expression for the difference in values between the globally optimum game and the locally optimum one can be easily written: a relative value $V_{jk}(x_1)$ may be defined:

$$V_{jk}(x_1) = \sum_{i=1}^{j} \Delta n f(x'_i, y'_i, z'_i) + \overset{\circ}{V}(x'_{j+1}) \tag{7}$$

then:

$$V_{jk}(x_1) = \sum_{1}^{j-k} \Delta n f(x'_i, y'_i, z'_i) + V(x'_{j-k+1}; k\Delta n) \tag{8}$$

and

$$|V(x_1; j\Delta n) - V_{jk}(x_1)| = \left|\Delta n \sum_{1}^{j-k} \lfloor f(x^*_i, y^*_i, z^*_i) - f(x'_i, y'_i, z'_i) \rfloor + V(x^*_{j-k+1}; k\Delta n) - V(x'_{j-k+1}; k\Delta n)\right| \tag{9}$$

Given a particular game solved from both global and local points of view, the above expression gives the exact difference in pay-offs. However, since our intention is to apply the local criterium without solving the global game, bounds for the difference could probably be derived in terms of $\overset{\circ}{P}(x)$ and values of games of duration up to $k\Delta n$ only.

It is easy to show that the difference must be bounded. [7] In fact, with very intuitive assumptions about f, g and the control vectors, a bound can be derived even if the k-games are not played optimally.

The Continuous Case

The continuous counterpart of the locally optimum criterion is a straightforward extension of the concepts used in the discrete model. It simplifies the boundary conditions, and, being a strictly forward method, seems to be ideal for on-the-spot real-time trajectory computations.

We use the notation and terminology of Isaacs [9] and formulate the game as follows. Let the globally optimum game be described by a set of n autonomous differential equations and let players I and II have control vectors ϕ and ψ respectively.

$$\dot{x} = f(x, \phi, \psi) \qquad x \epsilon E^n. \tag{10}$$

Let the optimal pay-off and the player intentions be described by the value function;

$$V(x) = \min_{\phi}\max_{\psi} \int G(x,\phi,\psi)dt + H(s) \;, \tag{11}$$

where the integration is performed along the trajectory in the state space, ξ, and the function $H(s)$ is the terminal pay-off, defined on a $n-1$ dimensional terminal (target) surface, $\mathscr{C}$, parameterized by the vector s.

We take the duration of the globally optimum game starting at x to be To, the time needed to reach $\mathscr{C}$ or the "stop rule" fixed time.

Consider a much shorter local game, of duration T<<To with value function $V(x,x_{n+1})$, where x is the state vector of the original system and x_{n+1} is a "clock variable". The first n+1 focal equations (written in retrogressive form) are given by

$$\begin{aligned} &\mathring{x}_i = -f_i(x,\phi,\psi) && i = 1,\ldots,n \\ &\mathring{x}_{n+1} = +1 && x_{n+1}(0) = 0 \qquad x_{n+1} = \tau \end{aligned} \tag{12}$$

where $\mathring{x}_i = \dfrac{dx_i}{d\tau}$, with τ being a parameter along the trajector emanating from the new terminus, $\mathscr{C}'=\xi\times\{0\}$, into the new space $\xi'=\xi\times[0,\infty)\subset E^{n+1}$. Isaacs first Main Equation, ME1, will be, in the zero sum case:

$$\min_{\phi}\max_{\psi} \sum_{1}^{n+1} V_i f_i(x,\phi,\psi) + G(x,\phi,\psi) = 0 \tag{13}$$

where $V_i = \dfrac{\partial V(x,x_{n+1})}{\partial x_i}$ and we are assuming that f and G do not depend on τ.

Since $V_{n+1} = \dfrac{\partial V}{\partial \tau}$ we write

$$\min_{\phi}\max_{\psi} \left[-\frac{\partial V}{\partial \tau} + \sum_{i=1}^{n} V_i f_i(x,\phi,\psi) \right] = 0 \tag{14}$$

From this we obtain $\bar{\phi}(x,V_x)$ and $\bar{\psi}(x,V_x)$, the optimizing tactics, where V_x denotes the gradient vector of the value function.

The second main equation becomes

$$-\frac{\partial V}{\partial \tau} + \sum_{1}^{n} V_i f_i(x,\bar{\phi},\bar{\psi}) + G(x,\bar{\phi},\bar{\psi}) = 0 \tag{15}$$

and the corresponding characteristic equations of this partial differential equation are given by

$$\begin{aligned} &\mathring{x}_i = -\, f_i(x,\bar{\phi},\bar{\psi}) && i = 1,\ldots,n \\ &\mathring{x}_{n+1} = +\,1 \\ &\mathring{V}_i = \sum_{j=1}^{n} V_i f_{ij}(x,\bar{\phi},\bar{\psi}) + G_i(x,\bar{\phi},\bar{\psi}) && i = 1,\ldots,n \\ &\mathring{V}_{n+1} = 0 \end{aligned} \tag{16}$$

where

$$f_{ij}(x,\bar{\phi},\bar{\psi}) = \frac{\partial f_i}{\partial x_j}\;(x,\bar{\phi},\bar{\psi}).$$

The Eulerian derivative of the value function is given by

$$\dot{V} = \sum_{1}^{n} V_i f_i(x,\bar{\phi},\bar{\psi}) - \frac{\partial V}{\partial \tau} = - G(x,\bar{\phi},\bar{\psi}) \ . \tag{17}$$

So the value will be constant along the path if and only if the pay-off is terminal. Hence, in the case of terminal pay-off the state travels on a surface of constant value.

Initial Conditions

The task of integrating the above system of equations remains unchanged. However, the boundary conditions are much simplified and make easier the final elimination of parameters and the obtaining of solutions.

The initial strip manifold is given by the n-dimensional manifold: $\mathcal{E}'=\xi\times\{0\}$. $\mathcal{E}'$ is parameterized by

$$\begin{aligned} x_i &= h_i(s_1,\dots,s_n) = s_i \quad i = 1,\dots,n \\ x_{n+1} &= 0 \end{aligned} \tag{18}$$

For termination points, the initial value of the value function (in the retrogressive sense) is given by a suitably chosen preference function defined on ξ and used in place of the terminal pay-off function of the original game, coinciding with its values on the former terminal surface.

$$V(x,x_{n+1}=0) = \overset{\circ}{V}(x) = \overset{\circ}{P}(x) = P(s_1,s_2,\dots,s_n). \tag{19}$$

Due to the simplicity of the parametric form of $\mathcal{E}'$, the evaluation of the initial values for the V_i is immediate:

$$\frac{\partial P^\circ}{\partial s_i} = \sum_{1}^{n} \frac{\partial V}{\partial x_k}\frac{\partial x_k}{\partial x_k} + \frac{\partial V}{\partial x_{n+1}} \cdot 0 = V_i \tag{20}$$

The last condition is given by ME 2 on $\mathcal{E}'$

$$\frac{\partial V}{\partial \tau}\Big|_{\mathcal{E}'} = V_{n+1}\Big|_{\mathcal{E}'} = \Big\lfloor \sum_{i=1}^{n} P_{si} f_i(s,\bar{\phi},\bar{\psi}) + G(s,\bar{\phi},\bar{\psi}) \Big\rfloor_{\mathcal{E}'} \tag{21}$$

The remaining solution proceeds in the usual form, with the locally optimum tactics being obtained by setting $x_{n+1} = T$.

The Dolicobrachistocrone

An example of the application of the locally optimum criterion to cases for which the global solution is known is given by the "Dolicobrachistocrone" problem of Isaacs. Its relation to the classical brachistocrone problem is obvious.

The problem consists in the motion of a point, free to move in the upper half of the x,y plane. Player I chooses the instantaneous direction of the velocity vector, the magnitude of which is $\sqrt{y}$. Player II superimposes another velocity vector from a linear vectogram. The terminal 1-dim. manifold is the positive y-axis and the playing space ξ is the first quadrant, as shown in Fig. 1.

The optimization is on time, with H(s)=0 and G=1. The K.E. are

$$x = \sqrt{y}\ \cos\phi + \frac{W}{2}(\psi+1)$$

$$y = \sqrt{y}\ \sin\phi + \frac{W}{2}(\psi-) \quad -1 \le \psi \le + 1 \tag{22}$$

The analysis shows a Bang-Bang type of control for II. The switching surface is given by:

$$A = V_x + V_y = 0 \tag{23}$$

The set of RPE is: $A > 0 \quad \bar{\psi} = +1$

$$\mathring{x} = \sqrt{y}\,\frac{V_x}{\rho} - W \qquad \mathring{V}_x = 0$$

$$\mathring{y} = \sqrt{y}\,\frac{V_y}{\rho} \qquad \mathring{V}_y = \frac{-\rho}{2\sqrt{y}} \tag{24}$$

where $\rho = \sqrt{V_x^2 + V_y^2}$ and with $\cos\bar{\phi} = -\frac{V_x}{\rho}, \sin\bar{\phi} = -\frac{V_y}{\rho}$ and for $A < 0$:

$$\mathring{x} = \sqrt{y}\,\frac{V_x}{\rho} \qquad \mathring{V}_x = 0$$

$$\mathring{y} = \sqrt{y}\,\frac{V_y}{\rho} + W \qquad \mathring{V}_y = \frac{-\rho}{2\sqrt{y}} \tag{25}$$

Integration yields:

$$y = \frac{s}{2}\left(1 + \cos\frac{\tau}{\sqrt{s}}\right) = y(s,\tau)$$

$$x = \frac{\tau}{2}\sqrt{s} + \frac{s}{2}\sin\frac{\tau}{\sqrt{s}} - W\tau = x(s,\tau) \tag{26}$$

until the first switching surface is given by

$$\tau = \tau_o = \frac{\sqrt{s}\,\pi}{2} \tag{27}$$

For our purposes, we will consider this portion of the solution only. Note that the above solution is left in parametric form.

As a Locally Optimum Game

The R.K.E. become:

$$x = -\left(\sqrt{y}\cos\phi + \frac{W}{2}(\psi+1)\right)$$

$$y = -\left(\sqrt{y}\sin\phi + \frac{W}{2}(\psi-1)\right)$$

$$z = +1 \quad z(0) = 0 \quad z = \tau \tag{28}$$

The ME1:

$$\min_{\phi}\max_{\psi} \left\lfloor V_x\left(\sqrt{y}\cos\phi + \frac{W}{2}(\psi+1)\right) + V_y\left(\sqrt{y}\sin\phi + \frac{W}{2}(\psi-1)\right) - V_z + 1 \right\rfloor = 0$$

or

$$-V_z + 1 + \min_{\phi}\max_{\psi} \left\lfloor \sqrt{y}\,(V_x\cos\phi + V_y\sin\phi) + \frac{W}{2}(V_x(\psi+1) + V_y(\psi-1)) \right\rfloor = 0 \tag{29}$$

Again:

$$\cos\bar{\phi} = \frac{-V_x}{\rho}, \sin\bar{\phi} = \frac{-V_y}{\rho}, \bar{\psi} = \operatorname{sgn} A, \; A = V_x + V_y \tag{30}$$

ME2:

$$F = -V_z - \sqrt{y}\,\rho + WV_x + 1 = 0 \qquad A > 0$$

$$F = -V_z - \sqrt{y}\,\rho - WV_y + 1 = 0 \qquad A < 0 \tag{31}$$

The characteristic equations of which are given by:

$$
\begin{aligned}
\mathring{x} &= \sqrt{y}\,\frac{V_x}{\rho} - w & \mathring{V}_x &= 0 \\
\mathring{y} &= \sqrt{y}\,\frac{V_y}{\rho} & \mathring{V}_y &= - /2\sqrt{y} \\
\mathring{z} &= +1 & \mathring{V}_z &= 0
\end{aligned}
\tag{32}
$$

Since $\mathscr{S}' = \xi \times \{0\}$, we let $x=s_1$, $y=s_2$ and $z=0$. For simplicity we choose the preference function $P(s_1,s_2)=s_1$. Application of (20) with ME2 on $\mathscr{S}'$ leads to $V_x=1$, $V_y=0$ and $V_z=- s_2 + w+1$ on $\mathscr{S}'$. One then finds $y=s_2/2\lfloor 1+\cos\tau/\sqrt{2}\rfloor$ and $\sin\phi= \sqrt{1-(y/s_2(y))}$. If we let $\tau=T$ we can find $s_2(y)$ and ϕ will depend only on y. Computed trajectories are shown in Fig. 2. The reader is asked to compare them with the globally optimum solutions in ⌊9⌋.

State Regulator Problem

We will now formulate the locally optimum theory for the state regulator problem of the following linear, time varying system:

$$x(t) = A(t)x(t) + B(t)u(t) \tag{33}$$

where $A(t)$ and $B(t)$ are system matrices of order $(n\times n)$ and $(n\times r)$ respectively. The pay-off is given by

$$J_1 = \frac{1}{2} \langle x(T_o), Fx(T_o)\rangle + \frac{1}{2}\int_{t_o=0}^{T_o} \lfloor \langle x(t),Q(t)x(t)\rangle + \langle u(t),R(t)u(t)\rangle \rfloor dt \tag{34}$$

where F is a constant $n\times n$ P.S.D. matrix, $Q(t)$ is an $n\times n$ P.S.D. matrix, and $R(t)$ is a P.D. matrix. A unique optimal control exists ⌊13⌋, and is given by

$$u(t) = -R^{-1}(t)B'(t)K(t)x(t) \tag{35}$$

where $K(t)$ is a $n\times n$ symmetric matrix, unique solution of the matrix Riccati Equation:

$$\dot{K}(t) = -K(t)A(t)-A'(t)K(t)+K(t)B(t)R^{-1}(t)B'(t)K(t)-Q(t) \tag{36}$$

with b.c. $K(T_o)=F$. $X(t)$ is then the solution of

$$\dot{x}(t) = \lfloor A(t)-B(t)R^{-1}(t)B'(t)K(t)\rfloor x(t) \qquad x(t_o) = x_o. \tag{37}$$

Now let us take a local game of duration $T<T_o$, and a pay-off given by

$$
\begin{aligned}
J_{1L} = \frac{1}{2}\langle x(t+T),\ Fx(t+T)\rangle + \frac{1}{2}\int_{t_o=T}^{t+T} &\lfloor \langle x(\tau),Q(\tau)x(\tau)\rangle + \\
&+ \langle u(\tau),R(\tau)u(\tau)\rangle\rfloor d\tau
\end{aligned}
\tag{38}
$$

for $t\le\tau\le t+T$. Again a unique optimal control exists and is given by

$$u(\tau) = -R^{-1}(\tau)B'(\tau)K^*(\tau)x(\tau) \qquad t\le\tau\le t+T$$

from the unique solution of

$$
\begin{aligned}
\dot{K}^*(\tau) &= -K^*(\tau)A(\tau) - A'(\tau)K^*(\tau) + K^*(\tau)B(\tau)R^{-1}(\tau)B'(\tau)K^*(\tau) - \\
&- Q(\tau)\ ;\ K^*(t+T) = F
\end{aligned}
\tag{39}
$$

To obtain the initial value of the optimal control, let $\tau=t$. Hence:

$$u(t) = -R^{-1}(t)B'(t)K^*(t)x(t) \qquad \text{replaces (35)} \tag{40}$$

Note that in (39) the values of A,B,R and Q are those of the original game for $t<\tau<t+T$. If these matrices are time-invariant, the solution of (39) will depend only on T, and $K^*(\tau=t)$ is a constant. Hence u(t) becomes proportional to the state until $t=T_0-T$. From there on, the last local game is played through, and K^* varies.

Example

$$x_1 = x_2 \quad F = 0 \quad J = \int_t^{t+\tau} (x_1^2+u^2)d\tau \tag{41}$$

$$x_2 = u \quad R = 1$$

the solution is given by [14]:

$$u(\tau') = - \frac{\cos^2 a\tau'\sinh^2 a\tau'+\sin^2 a\tau'\cosh^2 a\tau'}{\cosh^2 a\tau'+\cos a\tau'} x_1 - \frac{2\,(\cosh a\tau'\sinh a\tau'-\sinh a\tau'\cos a\tau')X_2}{\cosh^2 a\tau'+\cos^2 a\tau'} \tag{42}$$

with $a=1/\sqrt{2}$ and $\tau'=t+t-\tau$. To obtain the initial value of $u(\tau')$, we let $\tau=t$. (42) becomes $u(T)=\alpha(T)x_1+\beta(T)x_2$, with α and β the evaluated coefficients in (42). In summary, the locally optimal control is given by:

$$u(t) = \alpha(T)x_1 + \beta(T)x_2 \qquad \text{for } 0\le t\le T_0-T \tag{43}$$

and by (42) for $T_0-T\le t\le T_0$ with $\tau'=T_0-t$.

The locally optimum strategy then applies a control which is a time-invariant linear combination of the state variables until the game termination is "within range". From then on, the game is played optimally.

Comments

The optimizing method suggested in this paper is a mixture of essentially local, forward search methods and retrogressive, globally optimum methods [10]. As such, it has characteristics of both. It contains more than heuristic methods, which are considered as intelligently directed search procedures [11], but yet it is not "error-controlled". We have the assurance, however, that, when the globally optimum problem is of fixed-time type, the local value tends to the global one as $T\to T_0$, and the method might be aaid asymptotically optimum.

On the other hand, in collision avoidance as in many other optimization problem, the game termination is either remote or not well defined. It follows that the optimum solution is either not well defined or physically impossible to determine. Hence, there is no sense in considering an absolute "error", but only in ordering (partially or linearly) suboptimal solutions. [12]

Applications other than in differential games should be about as many as there are distinct optimization problems, both discrete and continuous. The statement of the problem in terms of "preferences" is very intuitive, and should find its way into many practical problems.

REFERENCES

1. Aris, Rutherford, Discrete Dynamic Programming. Blaisdell Publishing Co., 1964.

2. Von Neumann and Morgenstern. Theory of Game and Economic Behavior. Princeton, 1944.

3. deGroot, Adriaan. Thought and Choice in Chess. Mouton & Co. The Hague, the Netherlands.

4. Herbert A. Simon and Allen Newell. Information Processing in Computer and Man. American Scientist, September, 1964.

5. Simon, H. A. and Simon, P. A. Trial and Error Search in Solving Difficult Problems: Evidence from the Game of Chess. Behavioral Science, 7, 425-429, October, 1962.

6. Fleming, W. H., The Convergence Problem for Differential Games. Journal Math. Anal. and Appl., Vol 3 (1961), pp. 102-116.

7. Souza, C. R. and Ciletti, M. D. Locally Optimum Games and Vehicular Collision. Univ. of Notre Dame Technical Report No. EE-6512, Sept., 1965.

8. Courant, R. and Hilbert, D. Methods of Mathematical Physics. Vol. II, Interscience Publishers, 1953.

9. Isaacs, Rufus. Differential Games. Center for Naval Analysis, John Wiley & Sons, 1965.

10. Wood, C. F. Review of Optimization Techniques. IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-1, No. 1, Nov., 1965.

11. Reiter, S. and Sherman, G. Discrete Optimizing. J. Soc. Indust. Appl. Math., Vol. 13, No. 3, September, 1965.

12. Leake, R. J. and Liu, R. Construction of Sub-Optimal Sequences. Dept. of Elec. Engrg., Univ. of Notre Dame, August, 1965.

13. Athans, M and Falb, Peter. Optimal Control. McGraw-Hill, 1966.

14. Leake, R. J. Analytical Controller Design and the Servomechanism Problem. TR64-4, E.E. Dept., Syracuse University.

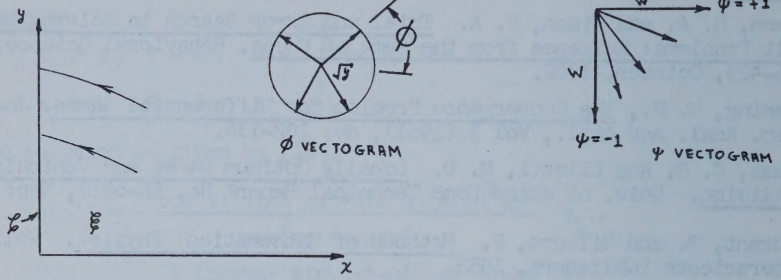

y
x
ℓ
ℓℓℓ
ϕ
$\sqrt{\gamma}$
ϕ VECTOGRAM
W
W
$\psi = +1$
$\psi = -1$
ψ VECTOGRAM

Figure 1

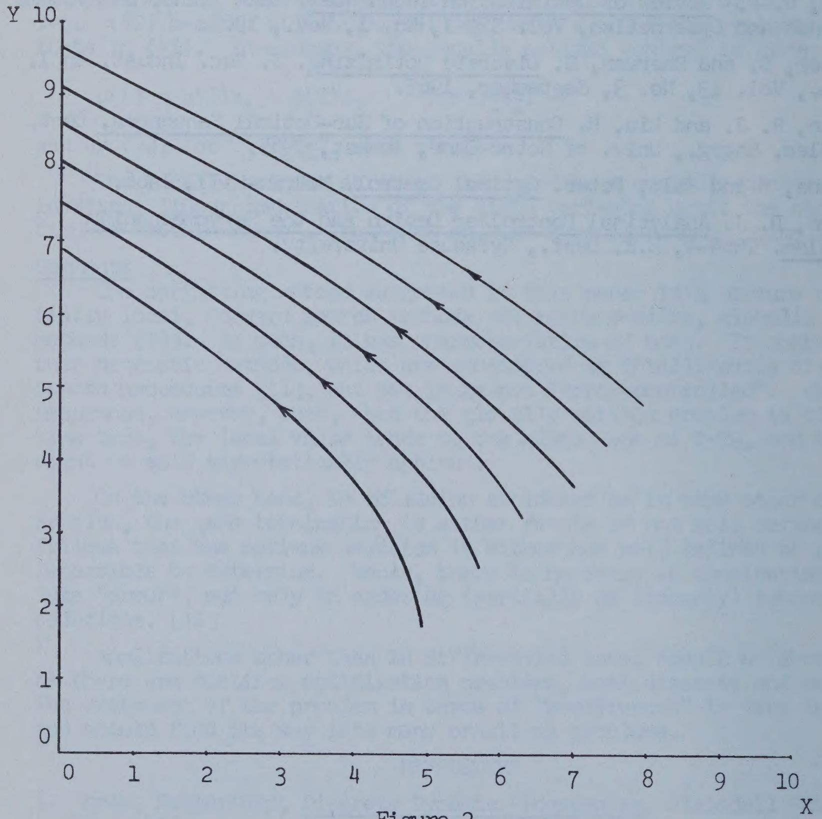


Figure 2

SIGNAL PROCESSING BY PASSIVE LINEAR NETWORKS

JAMES A. RESH
Michigan State University
East Lansing, Michigan

ABSTRACT

The paper provides a direct algebraic characterization of the signal processing capabilities of ordinary passive networks. The central contribution is a factorization theorem for modified Hurwitzian matrices; this leads to a minor new result on the formal realizability of state model specifications.

INTRODUCTION

We are concerned with the fundamental capabilities and limitations of passive lumped linear stationary networks, not as media for energy transfer, but as pure signal processors. Input signals are to be introduced by voltage and current sources, output signals are to be network voltages and currents, and the basic question is: How does passivity of the network limit the available input-output relations?

Our starting point is the now familiar fact that the input-output relation of any real lumped linear stationary system can be implicitly described by hybrid state equations [1] of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + \sum_{r=0}^{R} D_r u^{(r)}(t) \end{aligned} \tag{1}$$

when all the coefficient matrices are real. We made this choice of model because of its total lack of prejudice in the matter of signal identification—any component of the input and output vectors can, a priori, be either a voltage or a current.

Two basic properties of passive networks are pertinent to our problem. First, their zero-input responses are bounded functions of time; in terms of Eq. (1) this means that their A-matrices are modified Hurwitzian.[1] Second, differentiable inputs result in continuous outputs; from this it follows that $D_r = 0$ for $r \geq 2$. The remainder of our paper is devoted to a direct proof that these two necessary conditions are also sufficient for the existence of a passive network realization of the input-output relation described by Eq. (1).

---

[1] A matrix is Hurwitzian if all its eigenvalues have negative real parts; modified Hurwitzian if all its eigenvalues have nonpositive real parts and all its imaginary eigenvalues are simple zeros of the minimal polynomial.

AN ALGEBRAIC THEOREM

The heart of our proof is the theorem: Any real Hurwitzian or modified Hurwitzian matrix A possesses at least one factorization $A = -\Gamma R$ where Γ is real, symmetric and positive definite and R is real and positive definite or positive semidefinite.

When A is (strictly) Hurwitzian the theorem can be proven as follows.[2] Let Q be any real, symmetric, negative definite matrix of the same order as A; then the equation $LA + A'L = Q$ has a unique solution L, and this solution is real, symmetric and positive definite. [2] Define the real matrix $R \triangleq -LA$; to see that R is positive definite let x be any real nonzero column vector and observe: $x'Rx = -x'LAx = -\frac{1}{2}x'(LA+A'L)x = -\frac{1}{2}x'Qx > 0$. The desired factorization follows with $\Gamma \triangleq L^{-1}$.

If A is only modified Hurwitzian then the equation $LA + A'L = Q$ may have no solution and the above proof breaks down. For this case we have developed the following detailed argument.

It is well known that, since A is a square matrix with real entries, there exists for each real $\epsilon > 0$ a nonsingular matrix P_ϵ, and a matrix J_ϵ whose main-diagonal elements are the eigenvalues of A, such that

$$A = P_\epsilon J_\epsilon P_\epsilon^{-1} \quad . \tag{2}$$

Further, no loss of generality is entailed by the assumption that J_ϵ and P_ϵ can be described as follows.[3] J_ϵ can be partitioned as

$$J_\epsilon = \begin{bmatrix} J_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & J_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & J_2^* & 0 & 0 & 0 \\ 0 & 0 & 0 & J_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & J_3^* & 0 \\ 0 & 0 & 0 & 0 & 0 & J_4 \end{bmatrix} \tag{3}$$

where, for $k = 1, 2, 3,$ and 4,

---

[2] This concise proof in the Hurwitzian case was supplied by M. Sobral, Jr.

[3] To allow for the possibility that A does not possess one or the other of the eigenvalue types allowed for in Eq. (3) we invoke the obvious ad hoc definitions of matrices with no rows or no column or both.

$$J_k = \begin{bmatrix} \lambda_{k1} & e_{k_1} & 0 & \cdot & \cdot & 0 \\ 0 & \lambda_{k2} & e_{k_2} & \cdot & & \cdot \\ \cdot & 0 & \lambda_{k3} & \cdot & & \cdot \\ & & & & \cdot & \\ \cdot & & \cdot & \lambda_{k,n_k-2} & e_{k,n_k-2} & 0 \\ \cdot & & & 0 & \lambda_{k,n_k-1} & e_{k,n_k-1} \\ 0 & \cdot & \cdot & \cdot & 0 & \lambda_{kn_k} \end{bmatrix}$$

Here the superscript (*) indicates complex conjugate. The main-diagonal entries $\lambda_{k\nu}$ are real and nonzero for $k = 1$ and $\nu = 1, \ldots, n_1$; complex with nonzero real and imaginary parts for $k = 2$ and $\nu = 1, \ldots, n_2$; imaginary and nonzero for $k = 3$ and $\nu = 1, \ldots, n_3$; and zero for $k = 4$ and $\nu = 1, \ldots, n_4$. The entries $e_{k\nu}$ $(\nu = 1, \ldots, n_k;\ k = 1, 2, 3, 4)$ adjacent to the main diagonal are real, each equalling one or the other of the numbers $0, \epsilon$. Finally, when P_ϵ is partitioned to conform with the partitioning of J_ϵ, it assumes the form

$$P_\epsilon = [P_1 P_2 P_2^* P_3 P_3^* P_4] \quad . \tag{4}$$

For our purposes we need to recall just one more property of this modified Jordan form of A: If $\lambda_{k\nu}$ is a simple zero of the minimal polynomial of A then the associated off-diagonal entry $e_{k\nu}$ of J_ϵ is zero. From this, and the fact that the imaginary eigenvalues of modified Hurwitzian matrices are simple, it follows that J_3 and J_3^* are diagonal matrices and that J_4 is a zero matrix.

Next, let T be the matrix

$$T = \begin{bmatrix} 1_{n_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1_{n_2} & -j1_{n_2} & 0 & 0 & 0 \\ 0 & 1_{n_2} & j1_{n_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1_{n_3} & -j1_{n_3} & 0 \\ 0 & 0 & 0 & 1_{n_3} & j1_{n_3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1_{n_4} \end{bmatrix} \tag{5}$$

where 1_{n_k} denotes the unit matrix of order n_k and j is the imaginary unit. It is easy to show that T is nonsingular, and that its inverse is

$$T^{-1} = \begin{bmatrix} 1_{n_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2}\,1_{n_2} & \frac{1}{2}\,1_{n_2} & 0 & 0 & 0 \\ 0 & j\frac{1}{2}\,1_{n_2} & -j\frac{1}{2}\,1_{n_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}\,1_{n_3} & \frac{1}{2}\,1_{n_3} & 0 \\ 0 & 0 & 0 & j\frac{1}{2}\,1_{n_3} & -j\frac{1}{2}\,1_{n_3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1_{n_4} \end{bmatrix} \tag{6}$$

Direct multiplication reveals that the matrix $P_\epsilon T$ is real; and it is nonsingular since both factors are nonsingular.

Next, again by direct multiplication, we find that

$$T^{-1}J_\epsilon T = \begin{bmatrix} J_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \operatorname{Re}\{J_2\} & \operatorname{Im}\{J_2\} & 0 & 0 & 0 \\ 0 & -\operatorname{Im}\{J_2\} & \operatorname{Re}\{J_2\} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \operatorname{Im}\{J_3\} & 0 \\ 0 & 0 & 0 & -\operatorname{Im}\{J_3\} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

Clearly, $T^{-1}J_\epsilon T$ is real. Now consider the quadratic form $v'(T^{-1}J_\epsilon T)v$; partitioning v as $v' = [v_1' v_{21}' v_{22}' v_{31}' v_{32}' v_4']'$ to conform with the partitioning of $T^{-1}J_\epsilon T$ in Eq. (7) we find:

$$v'(T^{-1}J_\epsilon T)v = [v_1' v_{21}' v_{22}'] \begin{bmatrix} J_1 & 0 & 0 \\ 0 & \operatorname{Re}\{J_2\} & 0 \\ 0 & 0 & \operatorname{Re}\{J_2\} \end{bmatrix} \begin{bmatrix} v_1 \\ v_{21} \\ v_{22} \end{bmatrix} ; \tag{8}$$

recalling that the diagonal entries of J_1 and $\operatorname{Re}\{J_2\}$ are all strictly negative and that their off-diagonal entries are either 0 or ϵ, we see that the quadratic form on the right in Eq. (8) can be made negative definite by choosing ϵ sufficiently small. We conclude that for sufficiently small ϵ, say $\epsilon = \epsilon_o$, the matrix $T^{-1}J_{\epsilon_o}T$ is negative semidefinite.

Finally, we define the matrices

$$\Gamma \triangleq (P_{\epsilon_o}T)(P_{\epsilon_o}T)' \tag{9a}$$

$$R \triangleq [(P_{\epsilon_o}T)^{-1}]'(-T^{-1}J_{\epsilon_o}T)[(P_{\epsilon_o}T)^{-1}] . \tag{9b}$$

Observe that Γ is real, symmetric and positive definite: it is the product of a real nonsingular matrix and its transpose. And R is real and positive semidefinite: it is related by a real nonsingular congruence to the negative of a real negative semidefinite matrix.

Direct multiplication verifies that $-\Gamma R = P_{\epsilon_o} J_{\epsilon_o} P^{-1}_{\epsilon_o} = A$, completing the proof of our basic algebraic lemma.

THE PASSIVE SYNTHESIS

As discussed in the introduction, the hybrid state equation describing any passive ordinary network have the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + D_o u(t) + D_1 \dot{u}(t) \end{aligned} \tag{10}$$

where all the coefficient matrices are real and A is modified Hurwitz. In this section we show the converse, i.e., that any such hybrid state equations describe some passive network.

The realization which proves this existence theorem is shown in Fig. 1, where the various subnetworks are defined as follows:

λ is composed of positive inductors and ideal transformers, and constrains its port variables by the equation[4]

$$\nu_L(t) = \Gamma^{-1} \frac{di_L(t)}{dt} \quad ; \tag{11}$$

ρ consists of positive resistors, ideal transformers and gyrators, and realizes the constraint[4]

$$\nu_R(t) = R\, i_R(t) \tag{12}$$

β is an ideal transformer network whose port relations are[4]

$$\begin{aligned} \nu_{B_2}(t) &= (\Gamma^{-1}B)\, \nu_{B_1}(t) \\ i_{B_1}(t) &= -(\Gamma^{-1}B)'\, i_{B_2}(t) \end{aligned} \tag{13}$$

ξ is another network of ideal transformers, described by

$$\begin{aligned} \nu_{S_1}(t) &= -C'\, \nu_{S_2}(t) \\ i_{S_2}(t) &= C\, i_{S_2}(t) \quad ; \end{aligned} \tag{14}$$

---

[4] The matrices Γ and R are those in the factorization $A = -\Gamma R$ whose existence was demonstrated in the previous section.

δ is composed entirely of gyrators, its port equations being

$$\begin{aligned} i_{D_2}(t) &= D_o \nu_{D_1}(t) \\ i_{D_1}(t) &= -D'_o\, \nu_{D_2}(t) \quad ; \quad \text{and} \end{aligned} \tag{15}$$

γ is a network of positive capacitors and ideal transformers described by port equations of the form

$$\begin{aligned} i_{C_1}(t) &= C_{11} \frac{d\nu_{C_1}(t)}{dt} + D'_1 \frac{d\nu_{C_2}(t)}{dt} \\ i_{C_2}(t) &= D_1 \frac{d\nu_{C_1}(t)}{dt} + C_{22} \frac{d\nu_{C_2}(t)}{dt} \quad , \end{aligned} \tag{16}$$

where C_{11} and C_{22} are real symmetric matrices chosen so that $\begin{bmatrix} C_{11} & D'_1 \\ D_1 & C_{22} \end{bmatrix}$ is positive semidefinite.

The realizations of these subnetworks are provided by standard techniques. [3] To see that the composite network in fact realizes Eq. (2) is a simple matter of combining the above equations describing the subnetworks with the Kirchhoff laws imposed by the interconnection to obtain:

$$\begin{aligned} \dot{x}(t) = \frac{di_L(t)}{dt} = \Gamma\, \nu_L(t) &= \Gamma[v_R(t) + v_{B_2}(t) - v_{C_1}(t)] \\ &= \Gamma[\,Ri_R(t) + \Gamma^{-1}Bv_{B_1}(t) + C'\, v_{C_2}(t)] \\ &= \Gamma[\,-Rx(t) + \Gamma^{-1}B\, u(t) + C' \cdot 0] \\ &= -\,\Gamma Rx(t) + \Gamma\Gamma^{-1}u(t) \\ &= Ax(t) + Bu(t) \quad ; \end{aligned} \tag{17}$$

and

$$\begin{aligned} y(t) = i_{S_2}(t) + i_{D_2}(t) + i_{C_2}(t) &= Ci_{S_1}(t) + D_o v_{D_1}(t) + D_1 \frac{dv_{C_1}(t)}{dt} \\ &\quad + C_{22} \frac{dv_{C_2}(t)}{dt} \\ &= Cx(t) + D_o u(t) + D_1 \frac{du(t)}{dt} \quad . \end{aligned} \tag{18}$$

DISCUSSION

This work provides a clean contrast between the signal processing capabilities of passive networks and those of active networks: The real hybrid state equations Eq. (1) can always be realized by active networks, e.g., a properly programmed analog computer; passive networks are restricted to the special class indicated by Eq. (2) with modified Hurwitzian A. It is interested to note that in the context of system simulation--the major source of pure signal processing specifications for networks--stable systems describable by the canonical state equations

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

and without doubt the most common specification; and that our work supplies a direct proof of the fact that one needs never resort to electronic analog computers to perform this simulation task--passive networks are theoretically sufficient.

Although the present paper has some novel aspects, it is true that its subject never leaves well traveled territory; the basic questions doesn't become really vexing until it forbids the use of gyrators. To the best of our knowledge the problem has not yet been attacked when gyrators are excluded but ideal transformers are admitted. In contrast, a considerable amount of effort is being devoted to it when both gyrators and ideal transformers are excluded. [4, 5, 6, 7] That work is somewhat different than ours in spirit, however, because it makes no effort to separate signal processing from energy processing. Technically, the different lies in the equation of signal identification: Previous work has required u(t) and y(t) to be complementary vectors of port variables and we do not.

REFERENCES

[1] L. A. Zadeh and C. A. Desoer, Linear System Theory: The State Space Approach, McGraw-Hill Book Company, New York, N. Y., 1963.

[2] R. Bellman, Introduction to Matrix Analysis, McGraw-Hill Book Company, New York, N. Y., 1960.

[3] D. Hazony, Elements of Network Synthesis, Reinhold Publishing Corporation, New York, N. Y., 1963.

[4] R. E. Kalman, "On a New Characterization of Linear Passive Systems," Proc. First Allerton Conf. on Circuit and System Theory, Urbana, Ill., 1963.

[5] D. Rauch, "On the Realization of Time Domain Models of Real Linear Bi-element Systems," Ph. D. Thesis, Michigan State University, 1963.

[6] A. Dervisoglu, "Realization of the A-matrix of Half-Degenerate RLC Networks," Ph. D. Thesis, University of Illinois, 1964.

[7] R. Yarlagadda and Y. Tokad, "On the Realization of RLC Networks from the State Model Characterizations," Proc. Ninth Midwest Symposium on Circuit Theory, Fort Collins, Colorado, 1965.

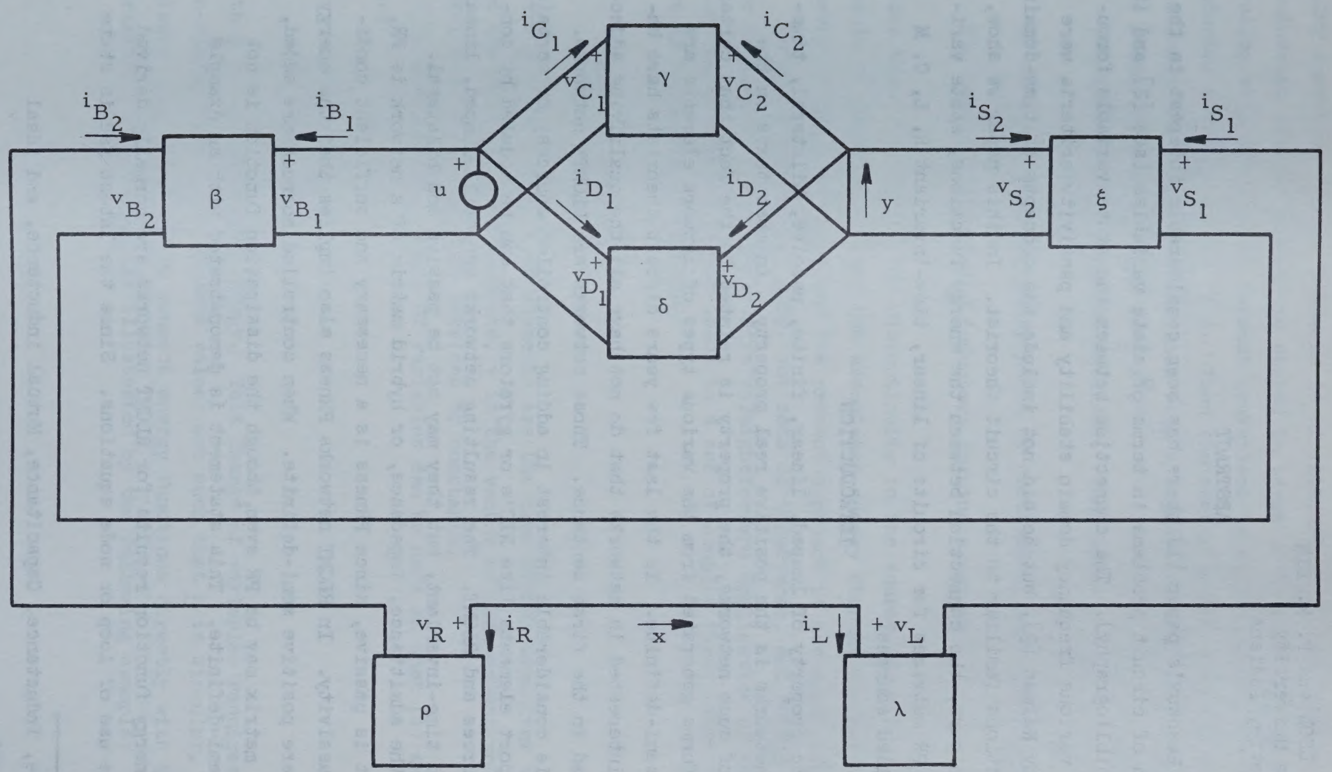

i_{B_2}
v_{B_2}
β
i_{B_1}
+
v_{B_1}
+
u
i_{C_1}
+
v_{C_1}
γ
+
v_{C_2}
i_{C_2}
i_{D_1}
+
v_{D_1}
δ
i_{D_2}
+
v_{D_2}
y
i_{S_2}
+
v_{S_2}
ξ
i_{S_1}
+
v_{S_1}
v_R
+
i_R
ρ
x
i_L
+
v_L
λ

Fig. 1. A Passive Realization.

STATE VARIABLES, ENERGY FUNCTIONS, AND PR MATRICES

B. J. LEON and P. M. LIN
Purdue University
Lafayette, Indiana

ABSTRACT

Since Bashkow's paper [1] there has been considerable interest in the formulation of circuit problems in terms of state variables [see [2] and its extensive bibliography]. The connection between the state variable formulation and various frequency domain stability and passivity criteria were discussed by Kalman [3], but he did not include the concepts of time-domain energy functions familiar to the circuit theorist. In this paper we show, at least in part, the connection between the energy functions, state variables, and PR matrices for circuits of linear, time-invariant R, L, C, M and controlled sources.

INTRODUCTION

A basic property of lumped, linear, finite, passive, bilateral, time-invariant networks is the positive real property. In most texts on the synthesis of such networks, the property is related to the fact that certain quadratic forms generated from the various types of network elements are positive, semi-definite. In the last few years circuit theorists have become more interested in networks that do not have all the qualifying adjectives listed in the first sentence. Those networks are RLCMT* networks. Now there is considerable interest in adding controlled sources, or certain other two port elements like NIC's or gyrators that can be modeled by controlled sources and RLCMT. The resulting networks are still lumped, linear, finite, and time-invariant, but they may not be passive and bilateral.

When the admittance, impedance, or hybrid matrix of a network is PR, the circuit is passive, since PRness is a necessary and sufficient condition for passivity. In RLCMT networks PRness also implies that the energy functions are positive semi-definite. When controlled sources are added, the hybrid matrix may be PR even though the dissipation function is not positive semi-definite. This statement is demonstrated with an example below.

The energy function results for RLCMT networks are normally derived through the use of loop or node equations. Since the latest fad is state variables, the relation between state variables and energy functions for RLCMT controlled source networks is of interest. For this reason the preliminaries to the example are stated in terms of state variables. When dealing with energies it is most convenient to choose state variable for inductor currents as $x = \sqrt{L}\, i$. Then the energy in the inductor is $\frac{1}{2} x^2$. Similarly the variable associated with a capacitor voltage is $x = \sqrt{c}\, v$. The stored electric energy is $\frac{1}{2} x^2$.

---

*Resistance, Inductance, Capacitance, Mutual inductance, and ideal Transformers

Two special cases must be excluded. They are:

1) Voltage sources, independent or controlled, and capacitors forming a loop;
2) Current sources and inductors forming a cut set.

Under these conditions any discontinuity in the source quantities will result in impulses through the associated energy storage elements. Then energy is not conserved, and the present approach is not applicable.

Loops of capacitance only, cut sets of inductance, and mutual inductors are included. For such circuits straightforward application of the topological formulation rules leads to a set of coupled first order differential equations in the form

$$\underset{\sim}{K}\, \dot{\underline{y}} = \underset{\sim}{F}\, \underline{y} + \underset{\sim}{G}\, \underline{v} \tag{1}$$

where $\underline{y}$ is a vector whose elements are the tree branch capacitor voltages and the link inductor currents, $\underline{v}$ is the vector of independent source inputs; and $\underset{\sim}{K}$, $\underset{\sim}{F}$, and $\underset{\sim}{G}$ are matrices of constants.

In most cases a change of variables can be made so that (1) assumes the form

$$\dot{\underline{x}} = \underset{\sim}{A}\, \underline{x} + \underset{\sim}{B}\, \underline{v} \tag{2}$$

with $\frac{1}{2} \underline{x}^T \underline{x}$ the stored energy. This change of variables requires $\underset{\sim}{K}$ to be non-singular. If the element values are such that $\underset{\sim}{K}$ is singular, the authors know of no way to connect energy functions directly with the state variables. The problem is illustrated by the following example:

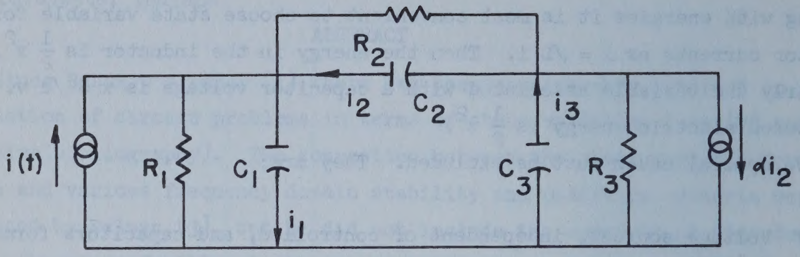


Figure 1

Consider the circuit of Figure 1. The arrows on the figure give current and voltage polarity with the passive sign convention. The current i_3 is through C_3 in the direction shown. Straightforward application of the topological rules for state variable formulation give

$$
\begin{aligned}
C_1 \dot{v}_1 &= -\frac{1}{R_1} v_1 - \frac{1}{R_3}(v_1 + v_2) + i_3 - \alpha C_2 \dot{v}_2 + i(t) \\
C_2 \dot{v}_2 &= -\frac{1}{R_2} v_2 - \frac{1}{R_3}(v_1 + v_2) + i_3 - \alpha C_2 \dot{v}_2
\end{aligned}
\tag{3}
$$

Since $i_3 = - C_3 (\dot{v}_1 + \dot{v}_2)$, (3) can be put in the form (1).

Thus

$$
\begin{aligned}
(C_1 + C_3)\dot{v}_1 + (C_3 + \alpha C_2)\dot{v}_2 &= -\left(\frac{1}{R_1} + \frac{1}{R_3}\right) v_1 - \frac{1}{R_3} v_2 + i(t) \\
C_3 \dot{v}_1 + (C_2(1 + \alpha) + C_3)\dot{v}_2 &= -\frac{1}{R_3} v_1 - \left(\frac{1}{R_2} + \frac{1}{R_3}\right) v_2
\end{aligned}
\tag{4}
$$

To relate the left side of (4) to the stored energy, we first define a change of variables

$$\underline{q} = \underset{\sim}{K}\, \underline{v} \tag{5}$$

where

$$\underset{\sim}{K} = \begin{bmatrix} C_1 + C_3 & C_3 + \alpha C_2 \\ C_3 & C_2 (1 + \alpha) + C_3 \end{bmatrix}$$

If $\underset{\sim}{K}^{-1}$ exists, then the stored energy W can be written in terms of the vector $\underline{q}$ as follows: First note that

$$W = \frac{1}{2} \underline{v}^T \underset{\sim}{C} \underline{v} \tag{6}$$

where

$$\underset{\sim}{C} = \begin{bmatrix} C_1 + C_3 & C_3 \\ C_3 & C_2 + C_3 \end{bmatrix}$$

Solving (5) for $\underline{v}$ in terms of $\underline{q}$ and substituting into (6) gives

$$W = \frac{1}{2} \underline{q}^T \underset{\sim}{S} \underline{q} \tag{7}$$

where

$$\underset{\sim}{S} = \underset{\sim}{K}^{-1T} \underset{\sim}{C} \underset{\sim}{K}^{-1}$$

Since $\underset{\sim}{C}$ is symmetric and positive semi-definite, $\underset{\sim}{S}$ is also. Thus $\underset{\sim}{S}$ can be written as

$$\underset{\sim}{S} = \underset{\sim}{F}^T \underset{\sim}{F}$$

where $\underset{\sim}{F}$ is non-singular. By defining a new variable

$$\underline{x} = \underset{\sim}{F} \underline{q}$$

the stored energy becomes

$$W = \frac{1}{2} \underline{x}^T \underline{x}$$

Now if $\underset{\sim}{K}$ is non-singular, (4) can be rewritten in terms of the variable $\underline{x}$

in the form (2). The result is

$$\underline{\dot{x}} = \underset{\sim}{F} \begin{bmatrix} -(\frac{1}{R_1} + \frac{1}{R_3}) & -\frac{1}{R_3} \\ -\frac{1}{R_3} & -\frac{1}{R_2} + \frac{1}{R_3}) \end{bmatrix} \underset{\sim}{K}^{-1} \underset{\sim}{F}^{-1} \underline{x} + [\begin{smallmatrix} i \\ 0 \end{smallmatrix}] \qquad (8)$$

PR MATRICES FROM STATE VARIABLES

To connect state variables, energy functions, and PR matrices, we start with a network characterized by the equations

$$\underline{\dot{x}} = \underset{\sim}{A}\, \underline{x} + \underset{\sim}{B}\, \underline{v}$$

$$\underline{y} = \underset{\sim}{C}\, \underline{x} + \underset{\sim}{D}\, \underline{v} + \underset{\sim}{E}\, \underline{\dot{v}} \qquad (9)$$

The state variables $\underline{x}$ are related to the total stored energy W by

$$W = \frac{1}{2} \underline{x}^{T} \underline{x} \qquad (10)$$

The vector $\underline{v}$ is an n-vector for an n-port system. Its elements are currents and voltages, one for each port. The $\underline{y}$ vector is also an n-vector of currents and voltages, one for each port. When an entry in $\underline{y}$ is a voltage the corresponding entry in $\underline{v}$ is a current and vice versa. Thus the input power is

$$P = \underline{v}^{T} \underline{y} \qquad (11)$$

The passive sign convention is taken at the ports. The dimension of the state vector $\underline{x}$ is determined by the number of independent points of energy storage. There is no definite relation between the dimension of $\underline{x}$ and the number of ports.

Since the dissipative energy P_d is the difference between the input power and the rate of change of stored energy, we may write

$$P_d = P - \dot{W}$$

$$= \underline{v}^T \underline{y} - \underline{x}^T \dot{\underline{x}}$$

$$= \underline{v}^T \underset{\sim}{C}\, \underline{x} + \underline{v}^T \underset{\sim}{D}\, \underline{v} + \underline{v}^T \underset{\sim}{E}\, \dot{\underline{v}} - \underline{x}^T \underset{\sim}{A}\, \underline{x} - \underline{x}^T \underset{\sim}{B}\, \underline{v} \qquad (12)$$

By Tellegem's Theorem P_d can also be computed as the power delivered to the resistors and controlled sources in the network.

It is well known that a passive network has a PR transfer function matrix, and that a network with a PR transfer-function matrix is passive. For an RLCMT network, PRness implies that the quadratic forms W and P_d are positive semi-definite and vice versa. When controlled sources are included in the network, positive semi-definiteness implies PRness, but not vice versa. The proof is as follows for networks characterized by (9) and (10).

The transfer function matrix H (s) can be generated from the normal form equations (9) by assuming a real time function input in the form

$$\underline{v}\,(t) = \underline{V}\, e^{st} + \underline{V}^{*}\, e^{s^{*}t} \qquad (13)$$

where $\underline{V}$ is a vector of complex numbers and the asterisk means complex conjugate. All other quantities in (9) have similar functional forms. That is,

$$\underline{x}\,(t) = \underline{X}\, e^{st} + \underline{X}^{*}\, e^{s^{*}t}$$

$$\underline{y}\,(t) = \underline{Y}\, e^{st} + \underline{Y}^{*}\, e^{s^{*}t} \qquad (14)$$

For each s the vectors $\underline{V}$ and $\underline{Y}$ are related by the transfer function matrix $\underset{\sim}{H}$ (s), as

$$\underline{Y} = \underset{\sim}{H}(s)\, \underline{V} \qquad (15)$$

Since the stored energy W is given by (10), the functional form (14) implies

$$W = \frac{1}{2}\left[\underline{X}^T \underline{X}\, e^{2st} + (\underline{X}^T \underline{X})^{*}\, e^{2s^{*}t} + 2\, |\underline{X}|^T\, |\underline{X}|\, e^{2\mathrm{Re}[s]t}\right] \qquad (16)$$

where $|\underline{X}|$ means a vector whose elements are the magnitude of the components of $\underline{X}$.

Similarly from (11) and (15)

$$P = (\underline{V}^{T} \underset{\sim}{H} \underline{V})\, e^{2st} + (\underline{V}^{T} \underset{\sim}{H} \underline{V})^{*}\, e^{2s^{*}t}$$

$$+ 2\mathrm{Re}\left[\underline{V}^{T*} \underset{\sim}{H} \underline{V}\right] e^{2\mathrm{Re}[s]t}$$

Then from (12)

$$\begin{aligned} P_d = e^{2\mathrm{Re}[s]t} \quad 2\Big\{ & \mathrm{Re}\,[\underline{V}^{T*} \underset{\sim}{H} \underline{V}] - \mathrm{Re}[s]\, |\underline{X}|^{T} |\underline{X}| \\ & + \left[\underline{V}^{T} \underset{\sim}{H} \underline{V} + \underline{s}\, \underline{X}^{T} \underline{X} \right] e^{2\mathrm{Im}[s]t} \\ & + \left[\underline{V}^{T} \underset{\sim}{H} \underline{V} + s\, \underline{X}^{T} \underline{X} \right]^{*} e^{-2j\mathrm{Im}[s]t} \Big\} \end{aligned} \tag{18}$$

If P_d is non-negative (positive semi-definite) for all real time functions, then the average value of the terms in the brace of (18) must be non-negative for all complex $\underline{V}$ and the corresponding $\underline{X}$ for every complex s. The first two terms in the brace are constants, while the last two together are a sinusoid of frequency 2 Im[s]. Thus the constant term must be at least as large as the peak value of the sinusoid.

Since $|\underline{X}|^{T} |\underline{X}|$ is positive for all complex $\underline{X}$, P_d positive semi-definite and Re $[s] \geq 0$ imply

$$\mathrm{Re}\left[\underline{V}^{T*} \underset{\sim}{H} \underline{V} \right] \geq \mathrm{Re}\,[s]\, |\underline{X}|^{T} |\underline{X}| \geq 0 \tag{19}$$

The inequality (19), along with the fact that $\underset{\sim}{H}(s)$ is real for real s, is the definition of a PR matrix. Thus P_d positive semi-definite implies $\underset{\sim}{H}(s)$ is PR.

To show that $\underset{\sim}{H}(s)$ PR does not imply the positive semi-definiteness of P_d, we consider the one port circuit of Figure 2.

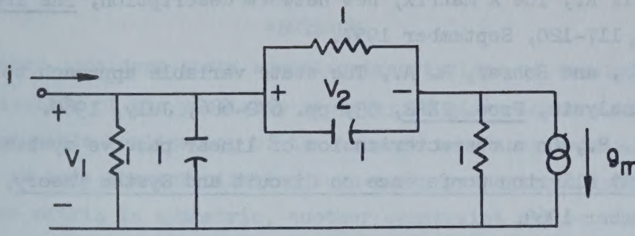


Figure 2

When i is the source at the port, the equations in the form (9) for this circuit are

$$\dot{v}_1 = -(2+g_m)\,v_1 + v_2 + i$$

$$\dot{v}_2 = (1+g_m)\,v_1 - 2\,v_2 \qquad (20)$$

$$v_1 = v_1$$

From (12)

$$P_d = (2+g_m)\,v_1^2 - (2+g_m)\,v_1 v_2 + 2v_2^2 \qquad (21)$$

For P_d to be positive semi-definite, we must have

$$-2 \le g_m \le 6 \qquad (22)$$

The input impedance is

$$Z(s) = \frac{s+2}{s^2 + (4+g_m)\,s + 3 + g_m} \qquad (23)$$

This impedance is PR so long as

$$-2 \le g_m < \infty \qquad (24)$$

Comparing (22) and (24) we see that (22) implies (24), but not vice versa.

BIBLIOGRAPHY

1. Bashkow, T. R., The A matrix, new network description, IRE Trans., CT-4, pp. 117-120, September 1957.
2. Kuh, E. S., and Rohrer, R. A., The state variable approach to network analysis, Proc. IEEE, 53, pp. 672-686, July, 1965.
3. Kalman, R. E., On a characterization of linear passive systems, Proc. First Allerton Conference on Circuit and System Theory, pp. 456-470, November 1963.

ACKNOWLEDGEMENT

This research was supported in part by the National Science Foundation, Grant GK26.

STATE REPRESENTATIONS, PASSIVITY, RECIPROCITY, AND N-PORT SYNTHESIS

DAVID M. LAYTON
TRW Systems, Redondo Beach, California

ABSTRACT

This paper considers state equation descriptions of passive networks. A characterization of the passivity property is given in terms of an arbitrary irreducible realization of the network impedance matrix, and the application of this characterization to n-port synthesis is indicated. If the impedance matrix is symmetric, another constraint may be added which enables the determination of a reciprocal network. These results become particularly simple for resistor-inductor networks.

1: INTRODUCTION

The application of state equation descriptions to linear systems has brought about a revival of interest in linear, passive network theory. This paper is a continuation of efforts in this direction. More specifically, the problems we consider here are the following:

(1) A positive real impedance matrix Z(s) is given; how may the passivity constraint be expressed in terms of an arbitrary, minimal set of state equations for the system described by Z(s)?

(2) If Z(s) is symmetric in addition to being positive real, how does this manifest itself in terms of the constraint of (1)?

(3) How may the above constraints be used to obtain passive network realizations for the given Z(s), both reciprocal and non-reciprocal?

Considerable work has been done on these and related problems by several investigators. An answer to (1) was given by Kalman for an impedance matrix with $Z(\infty) = 0$ in (1), and for a general driving point impedance in (2). A complete answer to (1) was given by Anderson in (3). Using the results of (3), Anderson and Newcomb gave a non-reciprocal solution to (3) above in (4). A complete solution to (3) was given in (5) by Youla & Tissi for the case when a scattering matrix rather than an impedance matrix is given. The characterization we give here is different from that of (3).

2: NOTATION AND PRELIMINARIES

The following notation will be used throughout:

- Z(s) is a positive real, rational, n x n matrix with $Z(\infty) < \infty$.
- the Mac Millan degree of Z(s) is denoted by δ (Z).
- the matrices $\{J, H, G, F\}$ will denote an arbitrary, real, constant, minimal realization of Z(s) (6); i.e., the system described by Z(s) is zero

state equivalent to the system described by:

$$\dot{\tilde{x}} = F\tilde{x} + G\tilde{u}$$

$$\tilde{y} = H^t\tilde{x} + J\tilde{u}$$

- capital letters will generally denote matrices, small letters with a tilde are vectors, and small letters with no tilde are scalars.

3: STATE EQUATION CHARACTERIZATION OF PASSIVITY AND NON-RECIPROCAL N-PORT SYNTHESIS

To characterize the passivity property of a given impedance matrix Z(s), we first use a result which has been demonstrated by a number of authors, namely, that there exists a passive network, possibly containing gyrators, for which Z(s) is the impedance matrix. Furthermore, this network contains exactly $\delta(Z)$ reactive elements.

Such a network is shown in Fig. 1; the reactances have been drawn external to a frequency independent subnetwork. No loss in generality is involved by assuming that each terminating reactance is a 1 Henry inductance, since ideal transformers and gyrators are permitted in the frequency independant network. The arguments which follow depend strongly on our ability to characterize the frequency independent network of Fig. 1, shown in the dotted lines. The following will suit this purpose:

Theorem I:

The frequency independent subnetwork of Fig. I always possesses an impedance description.

- The proof of Theorem I is fairly straightforward, and depends critically on the fact that exactly $\delta(Z)$ reactive elements are used in the network realization, which number is equal to the dimension of an irreducible set of state equations for Z(s) (6). The fact that Z(s) is regular at infinity is also significant.

Denote the impedance matrix of the frequency independent subnetwork by:

$$Z_1 = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$$

z_{11} is n x n, and z_{22} is $\delta(Z) \times \delta(Z)$. A straightforward analysis then gives:

$$Z(s) = z_{11} - z_{12}\,[sI + z_{22}]^{-1}\,z_{21}$$

We observe from this that the quadruple $\{z_{11}, z_{12}^t, -z_{21}, -z_{22}\}$ constitutes an irreducible realization of Z(s). Thus, if $\{J, H, G, F\}$ is any other minimal realization, there exists a constant non-singular matrix T such that

$$\begin{bmatrix} I & 0 \\ 0 & T^{-1} \end{bmatrix} \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix} = Z_1 \qquad (6)$$

i.e., a constant similarity transformation on the matrix

$$K = \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix}$$

yields a matrix which has positive semi-definite (p.s.d.) symmetric part.

We may obtain directly from this observation the desired passivity characterization. For a driving point impedance z(s), the theorem is as follows:

Theorem II:

Let z(s) be a rational function with $z(\infty) = j < \infty$, and $\{j, \underset{\sim}{h}, \underset{\sim}{g}, F\}$ an arbitrary real minimal realization. Then z(s) is positive real (p.r.) i.f.f. there exists a positive definite (p.d.) symmetric matrix $\underline{P}$ and a matrix L such that

$$\underline{P}F + F^t\underline{P} + \frac{1}{2j}\,(\underline{P}\,\underset{\sim}{g} - \underset{\sim}{h})\,(\underline{P}\,\underset{\sim}{g} - \underset{\sim}{h})^t = -L\,L^t \quad \ldots.. \qquad (*)$$

Outline of Proof

If z(s) is p.r. and $\{j, \underset{\sim}{h}, \underset{\sim}{g}, F\}$ is an arbitrary minimal realization, then from the remark stated above, there exists a non-singular matrix T such that the matrix

$$\begin{bmatrix} 1 & \underset{\sim}{0}^t \\ \underset{\sim}{0} & T^{-1} \end{bmatrix} \begin{bmatrix} j & \underset{\sim}{h}^t \\ -\underset{\sim}{g} & -F \end{bmatrix} \begin{bmatrix} 1 & \underset{\sim}{0}^t \\ \underset{\sim}{0} & T \end{bmatrix}$$

has p.s.d. symmetric part. Using this and letting $\underline{P} = T^{-1^t}\,T^{-1}$, we get after some manipulations:

$$\underline{P}F + F^t\underline{P} + \frac{1}{2j}\,(\underline{P}\,\underset{\sim}{g} - \underset{\sim}{h})\,(\underline{P}\,\underset{\sim}{g} - \underset{\sim}{h})^t$$

is n.s.d. Since this matrix is obviously symmetric, (*) follows directly.

If (*) holds for an arbitrary minimal realization, we perform any factorization of $\underline{P}$ or $\underline{P} = T^{-1^t}\,T^{-1}$. The matrix

$$\begin{bmatrix} 1 & \underset{\sim}{0}^t \\ \underset{\sim}{0} & T^{-1} \end{bmatrix} \begin{bmatrix} j & \underset{\sim}{h}^t \\ -\underset{\sim}{g} & -F \end{bmatrix} \begin{bmatrix} 1 & \underset{\sim}{0}^t \\ \underset{\sim}{0} & T \end{bmatrix}$$

may then be shown to have p.s.d. symmetric part, and thus may be realized as the series connection of a (passive) resistor - transformer network, and a gyrator network. Terminating appropriate ports of this frequency independent network with 1 H. inductances gives a passive network realization of z(s). Hence, z(s) is p.r.

Remark 1: To obtain the result of (1) (for the single input-output case, we observe that as $j \rightarrow 0$, (*) can be satisfied only if the p.s.d. term $(\underline{P}\,\underset{\sim}{g}-\underset{\sim}{h})(\underline{P}\,\underset{\sim}{g}-\underset{\sim}{h})^t$ is identically zero, i.e., if $\underline{P}\,\underset{\sim}{g} = \underset{\sim}{h}$. We thus have: $z(s)$ with $z(\infty) = 0$ is p.r. i.f.f.:

1) $\underline{P}F + F^t\underline{P}$ is n.s.d.
2) $\underline{P}\,\underset{\sim}{g} = \underset{\sim}{h}$

The result of (2) states that $z(s)$ is p.r. i.f.f. there exists a vector $\underset{\sim}{g}$ and a p.d. symmetric matrix $\underline{P}'$ such that:

$$\underline{P}'\,F + F^t\,\underline{P}' = -\underset{\sim}{g}\,\underset{\sim}{g}^t$$

$$\underline{P}'\,\underset{\sim}{g} - \underset{\sim}{h} = \sqrt{2j}\;\underset{\sim}{g}$$

We observe that this $\underline{P}'$ satisfies (*) with $L = 0$.

Remark 2: A result similar to Theorem II exists for the N-Port Case; this is given in (8).

Remark 3: The proof of the above theorem differs from (2) and (4) in that these authors gave constructive techniques for determining $\underline{P}$, whereas the above proof is of an existance nature. However, it does indicate explicitly how a passive network realization follows directly once the characterizing matrix $\underline{P}$ is found.

4: STATE EQUATIONS, PASSIVITY AND RECIPROCITY

We now consider impedance matrices $Z(s)$ which, in addition to being p.r., are also symmetric. We once again remove reactances from a minimal network realization, as was done previously; however, we now observe that the only gyrators in the network are those which appear where gyrator-inductor combinations were substituted for capacitors. The situation is indicated in Fig. 2. The frequency independent portion in the dotted lines has an impedance description, by Theorem 1, so that the open circuit (voltage) response is well defined at all terminal pairs, for current source excitation. Since the gyrator-current source combinations may be replaced by voltage sources, we have the following:

Theorem III:

The reciprocal, frequency independent subnetwork of Fig. 2 possesses a hybrid matrix description. The open circuit parts are the input ports and inductively terminated terminal pairs; the short circuit ports are the capacitively terminated terminal pairs.

Suppose the particular network realization we are dealing with contains k_1 inductances, and k_2 capacitances ($k_1 + k_2 = \delta(Z)$). Denote the hybrid matrix whose existence is claimed above by

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} ;$$

h_{11} is square of dimension $n + k_1$ and is the open circuit impedance matrix of the open circuit ports; h_{22} is square of dimension k_2, and is the short circuit admittance matrix of the short circuit ports. Then, since the frequency independent network is reciprocal, h_{11} and h_{22} are symmetric, and $h_{12} = -h_{21}^t$. Since the network is passive, h_{11} and h_{22} are p.s.d. Thus, H may be written:

$$H = \begin{bmatrix} h_{11} & h_{12} \\ -h_{12}^t & h_{22} \end{bmatrix}$$

with h_{11} and h_{22} symmetric and p.s.d.

If now we let

$$\Sigma = \begin{bmatrix} I_{nxn} & 0 & 0 \\ 0 & I_{k_1xk_1} & 0 \\ 0 & 0 & -I_{k_2xk_2} \end{bmatrix}$$

$$= \begin{bmatrix} I_{nxn} & 0 \\ 0 & \Sigma' \end{bmatrix} , \quad \Sigma' = \begin{bmatrix} I_{k_1xk_1} & 0 \\ 0 & -I_{k_2xk_2} \end{bmatrix} ,$$

then ΣH is symmetric. Furthermore, because the network realization employs the minimum number of reactive elements, if $\{J, H, G, F\}$ is any minimal realization of Z(s), then there exists a constant non-singular matrix T and a matrix Σ' of the form defined above such that the matrix:

$$\begin{bmatrix} I_{nxn} & 0 \\ 0 & \Sigma' \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T^{-1} \end{bmatrix} \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix}$$

is symmetric. Letting $A = T^{-1t} \Sigma' T^{-1}$, we obtain directly that the matrix:

$$\begin{bmatrix} I_{nxn} & 0 \\ 0 & A \end{bmatrix} \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix}$$

is symmetric.

Remark 1: Since $A = T^{-1^t} \Sigma' T^{-1}$, then A is symmetric and non-singular.

Remark 2: It may be shown (8) that A is uniquely determined by the (minimal) realization $\{J, H, G, F\}$.

Remark 3: Given the realization $\{J, H, G, F\}$, the matrix A is easily determined as the (unique) symmetric solution of the linear equations:

$$AG = -H$$
$$AF = F^t A$$

Remark 4: The number of +1's in the matrix Σ' correspond to inductances, the number of -1's to capacitances, in a minimal network realization of the impedance matrix Z(s). Since A is uniquely determined by $\{J, H, G, F\}$ and Σ' is uniquely determined from A (to within a permutation of its diagonal elements), we conclude that the number of capacitors and inductors is the same in any minimal network realization of Z(s). This agrees with a similar result for scattering matrices, shown in (5).

Remark 5: Once the matrix A is determined, any reduction to squares of the quadratic form $< \underset{\sim}{x}, A \underset{\sim}{x} >$ defines a transformation T from which a reciprocal, possible non-passive network realization may be obtained. Some subset of the networks so generated will be passive.

Remark 6: In the proof of the existence of A, the fact that minimal reactive passive reciprocal network realization for Z(s) exist was used; it can be shown (8) that the existence, uniqueness, non-singularity, and symmetry of A follow directly from only the symmetry of Z(s).

Let us proceed now to the primary purpose of this section, to express the reciprocity constraint in terms of the passivity constraint of Section 3. First of all, we determine a minimal realization of Z(s), such that

$$\begin{bmatrix} I & 0 \\ 0 & \Sigma' \end{bmatrix} \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix}, \Sigma' = \begin{bmatrix} I_{k_1} & 0 \\ 0 & -I_{k_2} \end{bmatrix},$$

is symmetric. This may easily be done, as described above. Suppose we now make a change of variables defined by T, and require that reciprocity be preserved, i.e., that the matrix

$$\begin{bmatrix} I & 0 \\ 0 & \Sigma' \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T^{-1} \end{bmatrix} \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix}$$

be symmetric. We obtain directly from the uniqueness of Σ' that T must satisfy:

$$T^{-1t}\,\Sigma'\,T^{-1} = \Sigma'$$

If in addition to preserving reciprocity, T also introduces passivity, then the matrix $\underline{P} = T^{-1t}\,T^{-1}$ satisfies a relation of the form (*) in Section 3. But, from above:

$$T^{-1t}\,\Sigma'\,T^{-1} = T^{-1}\,\Sigma'\,T^{-1t} = \Sigma'$$

$$\therefore T^{-1t}\,T^{-1}\,\Sigma'\,T^{-1t}\,T^{-1} = T^{-1t}\,\Sigma'\,T^{-1} = \Sigma'$$

$$= \underline{P}\,\Sigma'\,\underline{P}$$

i.e., the p.d. symmetric matrix $\underline{P}$ of (*) satisfies

$$\underline{P}\,\Sigma'\,\underline{P} = \Sigma' \quad \text{..........} \qquad (**)$$

(*) and (**) together thus give state equation constraints imposed by passivity and reciprocity.

The situation is considerably simplified if the only reactances present are capacitances or inductances. We will briefly consider this case in the next section.

5: RC, RL MINIMAL RECIPROCAL NETWORKS

We are given a positive real symmetric impedance matrix which has a passive, reciprocal network realization whose only reactive elements are inductances (or capacitances). We obtain a minimal realization $\{J, H, G, F\}$ such that, for some Σ',

$$\begin{bmatrix} I & 0 \\ 0 & \Sigma' \end{bmatrix} \begin{bmatrix} J & H^t \\ -G & -F \end{bmatrix}$$

is symmetric.

We make the following observations:

(1) Since all reactive elements are capacitances or inductances, $\Sigma' = \pm I$.

(2) Changes of variables preserving reciprocity satisfy

$$T^{-1t}\,\Sigma'\,T^{-1} = \Sigma',$$

which for the RC, RL case becomes:

$$T^{-1t}\,T^{-1} = I.$$

i.e., all such changes of variables are given by orthogonal matrices.

We now state the following:

Theorem IV:

All minimal reciprocal network realizations of RC, RL impedance matrices are passive.

Outline of Proof:

All minimal reciprocal network realizations of RL impedance matrices are generated by orthogonal transformations applied to an arbitrary minimal realization. The coordinate change is thus a congruence transformation. Since some of the network realizations are passive, then all must be, since the congruence transformation preserves passivity. A similar argument verifies the theorem for RC networks.

6: CONCLUDING REMARKS

Characterizations for passivity and reciprocity of positive real matrices in terms of minimal state decompositions of the matrix have been given which are different from those encountered in the literature. It should be emphasized that the proofs are of an existence nature; however, once the matrices whose existence is asserted are determined, a network synthesis, either reciprocal or non-reciprocal, may be straightforwardly accomplished.

ACKNOWLEDGEMENT

This research was conducted at the Electronics Research Laboratory, University of California, Berkeley, and supported in part by Joint Services Electronics Program Grant AF-AFOSR-139-66.

Also, the author would like to thank Prof. Ernest Kuh for numerous discussions concerning the research presented here and for reading this manuscript.

REFERENCES

(1) Kalman, R.E. - "A New Characterization of Linear Passive Systems" - Proc., 1st Allerton Conf.

(2) Kalman, R.E. - "Liapounov Functions for the Problem of Lure" - Proc. N.A.S., Vol. 49, 1963

(3) Anderson, B.D. - "A Control Theory Viewpoint of Positive Reality" - S.E.L. Report, February, 1966.

(4) Anderson, B.D., and Newcomb, R.W. - "Impedance Synthesis via State Space Techniques" - S.E.L. Tech. Rep. No. 6558-5.

(5) Youla, D.C., and Tissi, P. - "N-Port Synthesis via Reactance Extraction - Part I" - P.I.B. M.R.I. - 1309-66, January 1966.

(6) Kalman, R.E. - "Irreducible Realizations and the Degree of a Matrix of Rational Functions" - S.I.A.M. J., Vol. 13, No. 2, 1965.

(7) Oono and Yasuura - "Synthesis of Finite 2n-Terminal Networks with Prescribed Scattering Matrices" - Mem. Kyushu University, Vol. 14, 2, 1954.

(8) Layton, D.M. - Ph.D. Dissertation, University of California, Berkeley, 1966.

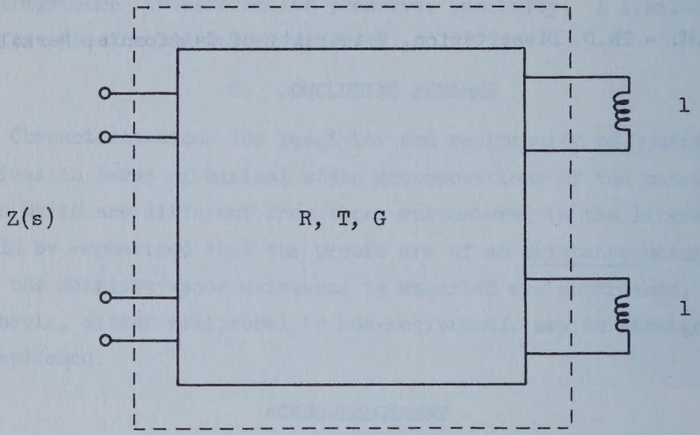


Fig. 1

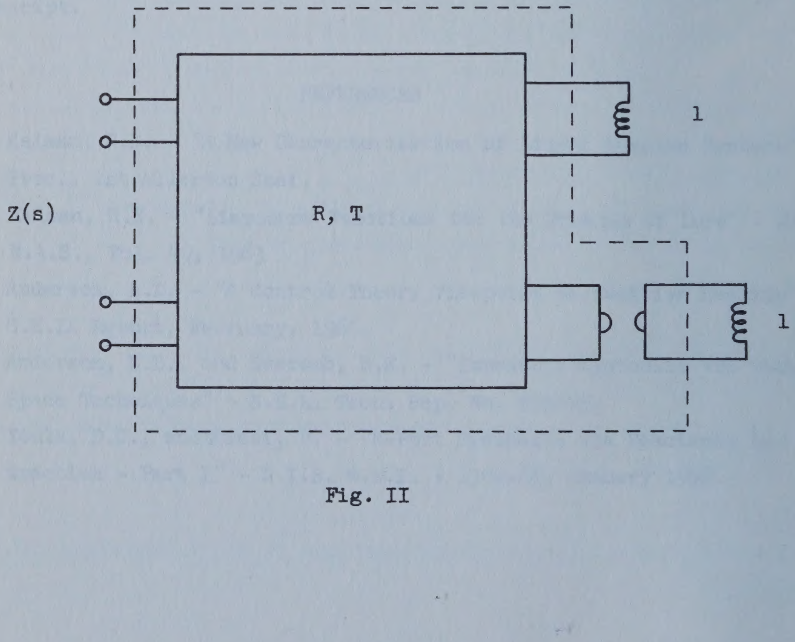


Fig. II

A CANONICAL NETWORK FOR PERIODICALLY VARIABLE LINEAR SYSTEMS

H. E. MEADOWS
Department of Electrical Engineering
Columbia University, New York, New York

L. M. SILVERMAN*
Department of Electrical Engineering
University of California, Berkeley, California

L. E. FRANKS **
Bell Telephone Laboratories, Inc.
North Andover, Massachusetts

ABSTRACT

A generalization of the n-path filter network is proposed for realization of the class of response functions associated with periodically variable linear systems. Like the original n-path filter, the proposed canonical network consists of a multiport fixed network connected between input and output modulators. The modulators may be regarded as linear time-variable controlled sources providing access to the fixed network. The fixed portion of the canonical network may be realized by RLC or RC elements.

The system to be realized may be described initially by an equivalent state-variable form, by a time-domain impulse response matrix, or by a time-variable system response matrix. The formal synthesis of the network from an impulse response matrix satisfying appropriate realizability conditions may be effected in the time domain to yield a minimal order realization. Realization from a periodic time-variable system function is also convenient; the realizability conditions in this case may be stated in terms of the consistency of a matrix product equation.

1. INTRODUCTION

In the theory of time-variable linear systems or networks, it has long been recognized [1-3] that periodically variable systems play a very significant role. From a practical viewpoint, the importance of their role in this theory appears quite reasonable, since many processes which are actually nonlinear may be studied effectively by perturbation methods leading to time-variable linear systems. If the strong signals present are periodic, the parametric variations presented to small signals are of course also periodic. Thus, for example, modulators and parametric amplifiers are frequently analyzed as periodically variable linear systems [4,5]. However, the potential uses of periodically variable networks and systems are far from exhausted, and in particular, their greater use in appropriate signal processing problems appears promising.

In this paper we shall develop basic synthesis procedures for realizing periodically variable linear systems for such purposes, using the network configuration indicated in Fig. 1. This configuration, which may be referred to as a generalized n-path network, has previously been employed in a considerably more specialized form to realize stationary system response functions [6]. Like the original n-path filter, the proposed canonical network consists of a multiport fixed network connected between input and output modulators. The modulators may be regarded as linear time-variable controlled sources providing access to the fixed network. The fixed portion of the generalized n-path network may be realized by RLC or RC elements. The latter possibility is especially desirable for low-frequency signal processing applications or for integrated circuits use.

---

*Formerly with the Department of Electrical Engineering, Columbia University.

**Associated with Columbia University during 1965 as faculty member in the Department of Electrical Engineering.

In view of the degree of practical success achieved with the earlier n-path filter for stationary system response synthesis, it appears that use of the more general network in Fig. 1 for nonstationary signal processing can be quite feasible. The applications envisioned here include real time operations not requiring a high-speed digital computer, for which other approaches might be more direct (as discussed in reference [7]). Only the realization problem is treated below; the approximation problem and other related topics will be discussed elsewhere.

Since the efforts of many investigators have been devoted to systems with periodic variations, we simply give here a reference to review articles for background material [8,9]. However, the closely related work in references [10-12] also should be noted.

We begin by showing in Section 2 that the generalized n-path network configuration provides a convenient and natural form for realizing the zero-state response of any periodically variable linear system. Synthesis from an equivalent state-variable description is then obvious. Formal synthesis in the time domain from an impulse response matrix satisfying appropriate realizability conditions is discussed in Section 3. Realization techniques using the frequency domain are developed in Section 4, where emphasis is given to system functions realizable with band-limited modulating functions.

2. CANONICAL NETWORK

The class of systems to be considered are those of the form

$$\dot{x}(t) = A(t)\,x(t) + \bar{B}(t)\,u(t) \tag{1a}$$

$$y(t) = \bar{C}(t)\,x(t) \tag{1b}$$

where $x(t)$, an n-vector, is the state of the system at time t; $\dot{x}(t)$ is its time derivative; $u(t)$, an r-vector, is the input; and $y(t)$, an m-vector is the output. The matrices $A(t)$, $\bar{B}(t)$, and $\bar{C}(t)$ are of order compatible with $x(t)$, $u(t)$, and $y(t)$ and will be assumed to be at least piecewise continuous. Furthermore it is supposed that the elements of each of the matrices has a common period which will be normalized to unity; i.e., $A(t) = A(t+1)$, $\bar{B}(t) = \bar{B}(t+1)$, $\bar{C}(t) = \bar{C}(t+1)$.

Let $X(t)$ be a fundamental matrix [13] for the homogeneous part of (1a), so that for all t,

$$\dot{X}(t) = A(t)\,X(t), \quad \det X(t) \neq 0. \tag{2}$$

As is well known, $X(t)$ may be expressed in the Floquet form [13]

$$X(t) = \bar{J}(t)\,e^{\bar{M}t} \tag{3}$$

where $\bar{M}$ is constant and $\bar{J}(t) = \bar{J}(t+1)$ is periodic.

Although $\bar{M}$ may not be a real matrix, it is always possible [13] to find real matrices M and J(t) such that

$$X(t) = J(t)\,e^{Mt} \tag{4}$$

and

$$J(t+2) = J(t) \tag{5}$$

Consider now the transformation of coordinates

$$z = J^{-1}x \tag{6}$$

which gives the equivalent system representation

$$\dot{z} = J^{-1}(AJ-\dot{J})z + J^{-1}\overline{B}\,u$$
$$y = \overline{C}Jz \qquad (7)$$

From (2) and (4), however, it is easily shown that

$$J^{-1}(AJ-\dot{J}) = M\,. \qquad (8)$$

Thus it is clear from the Floquet theory that any periodic system may be represented as a fixed system with periodic input and output multipliers, as in the network configuration of Fig. 1.

For simplicity of presentation we will restrict the following discussion to the case where M has n simple eigenvalues. When this is true, it is always possible to find a real nonsingular constant matrix T such that

$$T^{-1}MT = \Lambda$$

in which Λ has the form

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}$$

where $\Lambda_1 = \text{diag}\,\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$
and Λ_2 is a direct sum of second-order matrices

$$\Lambda_2 = \begin{bmatrix} \begin{bmatrix} \alpha_{k+1} & -\beta_{k+2} \\ \beta_{k+2} & \alpha_{k+1} \end{bmatrix} & & 0 \\ & \ddots & \\ 0 & & \begin{bmatrix} \alpha_{n-1} & -\beta_n \\ \beta_n & \alpha_{n-1} \end{bmatrix} \end{bmatrix}$$

The first k eigenvalues of M are real, $\lambda_1, \lambda_2, \ldots, \lambda_k$, while the complex eigenvalues are $\lambda_i = \alpha_i + j\,\beta_{i+1}$ and $\lambda_{i+1} = \alpha_i - j\,\beta_{i+1}$ for $i = k+1, \ldots, n$.

Thus if we perform the transformation of coordinates

$$w = T^{-1}x \qquad (10)$$

we obtain the equivalent system representation

$$\dot{w} = \Lambda w + B(t)u$$
$$y = C(t)w \qquad (11)$$

where Λ is given in (9), and B(t) and C(t) are periodic matrices with period 2 given by

$$B(t) = (J(t)T)^{-1}\,\overline{B}(t)$$
$$C(t) = \overline{C}(t)\,(J(t)T) \qquad (12)$$

The n-path realization of such a system could consist of simple RC and RLC networks with periodic controlled sources, all of the same period.

In the single-input single-output case, one possible realization is shown by Figs. 2-4. The subnetworks η_i for i = 1, 2, ..., k shown in Fig. 3 satisfy the RC network state-variable equations

$$\dot{q}_i = -q_i/\tau_i + b_i(t)\,u$$
$$y_i = c_i(t)\,q_i \qquad (13)$$

while those for i = k+1, k+3, ..., n-1 shown in Fig. 4 satisfy the RLC network state-variable equations

$$\begin{bmatrix} \dot{q}_i \\ \dot{\phi}_i \end{bmatrix} = \begin{bmatrix} 0 & 1/L_i \\ -1/C_i & -R_i/L_i \end{bmatrix} \begin{bmatrix} q_i \\ \phi_i \end{bmatrix} + \begin{bmatrix} b_i(t) \\ b_{i+1}(t) \end{bmatrix} u \qquad (14a)$$

and

$$y_i = \begin{bmatrix} c_i(t) & c_{i+1}(t) \end{bmatrix} \begin{bmatrix} q_i \\ \phi_i \end{bmatrix} \qquad (14b)$$

This realization of a pair of complex conjugate poles has a minimum dynamic order (no redundant reactors). It is therefore considerably simpler than a (nonminimal) second-order realization suggested previously by Sandberg [14].

Because the presence of inductors may be undesirable or impractical in low-frequency applications or integrated circuits use, it is important to observe the conditions under which only RC networks are required in the fixed subnetwork. These conditions, involving the eigenvalues of M, may be explored by lifting the restriction that all time variations have the same period. We need only consider the problem of replacing a single complex pole pair subnetwork of Fig. 2 by a subnetwork realizing real poles, i.e., having an RC fixed network. Let the equations describing the complex pole pair network be written as in (1) with

$$A = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}$$

Then we can exhibit a periodic transformation of coordinates T(t) such that (1) is (algebraically) equivalent to a new system with A replaced by

$$\bar{A} = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$$

Specifically, if we let $z = T^{-1}x$ to obtain

$$\bar{A} = T^{-1} AT - T^{-1} \dot{T}$$

then T must be a solution of the differential equation

$$\dot{T} = AT - T\bar{A} \qquad (15)$$

The general solution of (15) is

$$T = e^{At}Ke^{-\bar{A}t} \qquad (16)$$

where K is an arbitrary nonsingular constant matrix. Since $\bar{A}$ is a scalar matrix,

$$T = e^{(A-\bar{A})t} K, \quad A-\bar{A} = \begin{bmatrix} 0 & -\beta \\ \beta & 0 \end{bmatrix} .$$

Thus the desired transformation is simply

$$T(t) = \begin{bmatrix} \sin \beta t & -\cos \beta t \\ \cos \beta t & \sin \beta t \end{bmatrix} \qquad (17)$$

and the new modulator functions will be products of functions with periods $2\pi/\beta_i$ and functions with period 1. Thus, unless all β_i are rational multiples of 2π, an exactly periodic set of modulating functions is not possible with an RC network. However, it may be feasible to use products of periodic modulating functions in a physical realization. Moreover, since we may approximate irrational numbers by rational numbers arbitrarily closely, it is possible to realize all periodic systems within physical tolerances using periodic multipliers and RC fixed networks. This completes the synthesis from a description in state-variable form.

3. FORMAL SYNTHESIS IN THE TIME DOMAIN

Having demonstrated that the generalized n-path network may be regarded as canonical for finite order periodically variable linear systems, we can now show a formal synthesis procedure beginning with the time-domain impulse response function $H(t,\tau)$, the response at time t to a unit impulse input incident at time τ. The essential conditions on $H(t,\tau)$ may be stated as follows:

Theorem. If

$$H(t,\tau) = H(t+1, \tau+1) \tag{18}$$

and $H(t,\tau)$ is separable in the form

$$H(t,\tau) = \Psi(t)\ \Theta(\tau) \tag{19}$$

where $\Psi(t)$ has n columns which are linearly independent functions on $(-\infty,\infty)$ and $\Theta(\tau)$ has n rows which are linearly independent functions on $(-\infty,\infty)$, then $H(t,\tau)$ is realized by a system of order n of the form

$$\begin{aligned} \dot{x} &= Mx + B(t)\ u \\ y &= C(t)\ x \end{aligned} \tag{20}$$

where M is a constant matrix and

$$B(t) = B(t+1),\ C(t) = C(t+1). \tag{21}$$

Moreover, the realization is of minimum order.

Proof. In a related paper [15], Youla shows that finite intervals Δ_1 and Δ_2 exist such that

$$M_1 = \int_{\Delta_1} \Theta(\tau)\Theta'(\tau)d\tau \ \text{ and } \ M_2 = \int_{\Delta_2} \Psi'(t)\Psi(t)dt$$

are nonsingular. Similarly, let

$$N_1 = \int_{\Delta_1} \Theta(\tau-1)\ \Theta'(\tau)d\tau \ \text{ and } \ N_2 = \int_{\Delta_2} \Psi'(t)\ \Psi(t+1)dt.$$

Now, if we observe that (18) and (19) imply

$$\Psi(t+1)\ \Theta(\tau) = \Psi(t)\Theta(\tau-1) \tag{22}$$

and post-multiply (22) by $\Theta'(\tau)$, we find after integrating with respect to τ over Δ_1 that

$$\Psi(t+1) = \Psi(t)\ T_1 \tag{23}$$

where $T_1 = N_1 M_1^{-1}$, since M_1 is nonsingular. Similarly,

$$\Theta(\tau-1) = T_2\ \Theta(\tau) \tag{24}$$

where $T_2 = M_2^{-1} N_2$. Substituting (23) and (24) in (22) yields

$$\Psi(t)\ T_1\ \Theta(\tau) = \Psi(t)\ T_2\ \Theta(\tau)$$

which implies, by similar reasoning, that $M_2 T_1 M_1 = M_2 T_2 M_1$ or, since M_1 and M_2 are nonsingular, that

$$T_1 = T_2 = T \tag{25}$$

Since the rows of $\Psi(t)$ are linearly independent functions over all time, those of $\Psi(t+1)$ are also. Therefore, (23) implies that T is nonsingular and that

$$\Psi(t) = \Psi_1(t)\ T^t, \quad \Psi_1(t) = \Psi_1(t+1),$$

or, since T is nonsingular

$$\Psi(t) = \Psi_1(t)\ e^{Mt}$$

where $e^M = T$. Similarly,

$$\Theta(\tau) = e^{-M\tau}\ \Theta_1(\tau),\ \Theta_1(\tau) = \Theta_1(\tau+1),$$

so that setting $B(t) = \Theta_1(t)$ and $C(t) = \Psi_1(t)$ completes the synthesis. If M is not real, the period may be doubled, as before, to obtain a real matrix M.

The minimality of the realization may be shown by contradiction, using the Cayley-Hamilton theorem. The proof is omitted for brevity. The separability requirement in (19) is discussed further in reference [16].

4. SYNTHESIS IN THE FREQUENCY DOMAIN

The beginning point for synthesis in the frequency domain is the time-variable system function [1]. This description of system response appears to be of considerable value in synthesis, even though it is essentially equivalent to a time-domain approach. If the system to be realized has impulse response $H(t,\tau)$, the corresponding time variable system function is

$$W(s,t) = \int_0^\infty e^{-s\xi}\, G(\xi,t)d\xi \tag{26}$$

where ξ is the "age variable," and

$$G(\xi,t) = H(t,t-\xi) . \tag{27}$$

Let us again denote the system to be synthesized by

$$\begin{aligned} \dot{x} &= Mx + B(t)\,u \\ y &= C(t)\,x \end{aligned} \tag{28}$$

where M is constant, and $B(t) = B(t+1)$, $C(t) = C(t+1)$. For simplicity of notation, a single-input single-output system will be considered in this section.

It was shown by Darlington [2] that the singularities of $W(s,t)$ in the s plane are poles with positions independent of t for any periodic system described by linear differential equations. In terms of (28), these poles are simply the eigenvalues λ_i of matrix M and their congruent points $\lambda_i + j2\pi k$, where k is any integer. Moreover, the residues of W in these poles are periodic functions, which may be expanded in Fourier series. Thus, the initial description of the system to be realized is

$$W(s,t) = \sum_{i=1}^{n} \sum_{k} \frac{W_{ik}(t)}{s - (\lambda_i + j2\pi k)} \tag{29}$$

where

$$W_{ik}(t) = W_{ik}(t+1).$$

If we expand $W_{ik}(t)$ in a Fourier series, the complete system is specified (for zero-state response) by the poles $\lambda_i + j2\pi k$ and the Fourier coefficients of their residues. It will be convenient to expedite the writing of matrix equations later if these Fourier coefficients are indexed in the following manner:

$$W(s,t) = \sum_{i=1}^{n} \sum_{k} \sum_{\ell} \frac{N^i_{\ell k} e^{j2\pi(k-\ell)t}}{s-\lambda_i - j2\pi k} \tag{30}$$

The Fourier coefficients are then specified by the matrices

$$N^i = \left[N^i_{\ell k} \right] \tag{31}$$

so that ℓ indexes rows of N^i and k indexes columns. The poles λ_i lying in the region $-\pi < \text{Im } s \leq \pi$ and the corresponding matrices N^i are to be regarded as describing the system to be synthesized. Note that the s-plane periodicity of the poles is necessary for exact synthesis. The matrices N^i might be of infinite order in general, but emphasis will be placed here on the band-limited case where $N^i_{\ell k} = 0$ for $|\ell|$ or $|k|$ large enough. This requirement is imposed here to avoid the complications of infinite-order matrices, but it is not without some practical justification.

The periodicity of poles of $W(s,t)$ in the complex frequency plane is restrictive only if the fixed subnetwork is to be RC. Since this case is of sufficient practical importance we direct attention to it and require that all λ_i be real and negative, for passive RC realization of the fixed subnetwork. Then the eigenvalues of M are well defined, and M must be given by (9) with $\Lambda = \text{diag}\,\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. The matrix T is an arbitrary nonsingular constant matrix in general, but $M = \Lambda$ may be imposed to obtain n isolated signal paths. However, it is important to note that isolation of the paths as in the original n-path filter is not necessary. Any RC network with appropriate pole locations might be used, provided that certain algebraic conditions mentioned later are satisfied.

To proceed to the realizability conditions, the modulator functions $B(t)$ and $C(t)$ may be expanded in Fourier series as

$$B(t) = \sum_{k} e^{j2\pi kt}\, B_k \tag{32}$$

and

$$C(t) = \sum_{\ell} e^{-j2\pi \ell t} C_{\ell} \,. \tag{33}$$

Note the "reversed" indexing of C_{ℓ} in (33). Now the impulse response of the system (28) is

$$H(t,\tau) = C(t)\, e^{(t-\tau)M} B(\tau) \,. \tag{34}$$

so that $G(\xi,t)$ and $W(s,t)$ are readily determined. It will be convenient in (34) to express (33) as

$$C(t) = \begin{bmatrix} e^{j2\pi rt} & \dots & 1 & \dots & e^{-j2\pi rt} \end{bmatrix} \begin{bmatrix} C_{-r} \\ \vdots \\ C_0 \\ \vdots \\ C_r \end{bmatrix} \tag{35}$$

and to define matrices

$$\beta = \begin{bmatrix} B_{-q} & \dots & B_0 & \dots & B_q \end{bmatrix}, \quad \gamma = \begin{bmatrix} C_{-r} \\ \vdots \\ C_0 \\ \vdots \\ C_r \end{bmatrix} \tag{36}$$

We have assumed, as mentioned earlier, that $B(t)$ and $C(t)$ are band limited. It follows immediately from (27), (32), and (34) that

$$G(\xi,t) = C(t) \sum_{k} e^{\xi M + j2\pi k(t+\xi)} B_k$$

and the system function may be written as

$$W(s,t) = C(t) \sum_{k} (sI-M - j2\pi kI)^{-1} B_k\, e^{j2\pi kt}$$

where I represents the unit matrix of order n. Let us assume that the partial fraction expansion of matrix $(sI-M)^{-1}$ has the simple form

$$(sI-M)^{-1} = \sum \frac{1}{s-\lambda_i} K_i \tag{37}$$

where K_i are residue matrices. This expansion is possible, for example, in the ordinary n-path filter network even though eigenvalues are repeated. In fact, if all n paths are isolated and identical single-pole RC networks, we may write only one diagonal term in (37). If the paths are isolated and not identical, (37) may be written as a direct sum of such matrices. In more general cases it may contain n terms.

Using (37), we may express $W(s,t)$ as

$$W(s,t) = C(t) \sum_{i} \sum_{k} \left(\frac{1}{s-\lambda_i - j2\pi k} \right) K_i\, B_k\, e^{j2\pi kt}$$

or, using (35) also, as

$$W(s,t) = \sum_{i} \sum_{k} \sum_{\ell} C_{\ell}\, K_i\, B_k\, \frac{e^{j2\pi(k-\ell)t}}{s-\lambda_i - j2\pi k} \tag{38}$$

Thus, by comparison with (30), we may identify

$$N^{i}_{\ell k} = C_{\ell}\, K_i\, B_k \,. \tag{39}$$

Moreover, we may express N^i as

$$N^{i} = \left[N^{i}_{\ell k} \right] = \gamma\, K_i\, \beta \tag{40}$$

This last equation is the essential relation to be satisfied in the synthesis process. Together with the periodic pole spacing in the s plane, it provides the realizability conditions in terms of simple algebraic conditions. The matrix elements $N^{i}_{\ell k}$ must also meet appropriate symmetry conditions in order for the modulator networks to be real. For the case where all λ_i are real, they must satisfy

$$N^{i}_{\ell k} = \left(N^{i}_{-\ell,-k} \right)^{*} \tag{41}$$

where * denotes the complex conjugate.

To clarify somewhat the nature of the matrices N^i, we may observe that if all N^i are diagonal matrices, the complete system is then equivalent to a fixed system. Synthesis of a fixed system response was also treated in [6]. The off-diagonal terms in N^i represent modulation products. To illustrate (40), let us consider the synthesis of a system function

$$W(s,t) = \frac{4 + (1-j)e^{-j4\pi t}}{s + 1 + j2\pi} + \frac{2}{s + 1} + \frac{4 + (1+j)e^{j4\pi t}}{2 + 1 - j2\pi}$$

The matrix N^i is

$$N = \begin{bmatrix} 4 & 0 & (1+j) \\ 0 & 2 & 0 \\ (1-j) & 0 & 4 \end{bmatrix}$$

Since N has rank 3, it is clear that matrix $K = K^i$ in (40), as well as γ and β, must have rank at least 3. Therefore we may let $M = I = K$ and choose β arbitrarily as

$$\beta = \begin{bmatrix} 1 & 1 & 1 \\ -j & 1 & j \\ -1 & 1 & -1 \end{bmatrix}$$

The resulting unique matrix γ from (40) is

$$\gamma = \frac{1}{2} \begin{bmatrix} (2-j1) & (1+j3) & (-3-j2) \\ 2 & 0 & 2 \\ (2+j1) & (1-j3) & (-3+j2) \end{bmatrix}$$

Of course, the conditions for B(t) and C(t) real are that $B_{-k} = B_k^*$ and $C_{-k} = C_k^*$, which requires that the symmetry conditions on entries in N must be observed.

Equation (40) may be employed in several ways depending upon physical requirements. For example, an RC network with isolated identical paths may be used with preset input modulators and adjustable output modulators to achieve an adjustable network for time-variable signal processing. The required algebraic condition from (40) is essentially the solvability of well-determined linear algebraic equations for matrix γ. The solvability of these equations is in effect, a condition on observability or, in the dual case, on controllability of the preset portion of the system.

We may conclude this section by observing that the linear algebraic relations derived above make the realization of periodic systems a straightforward process that is computationally practical. The physical implementation of the canonical network is as feasible in the non-stationary case as in the previous work on n-path filters. It is expected that the approximation problem for signal processing and the use of pulse and step modulation in realization will be treated elsewhere.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation under grant NSF GP-2789 and the Office of Naval Research under contract NONR 4259(04). Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

REFERENCES

1. L.A. Zadeh, "Frequency Analysis of Variable Networks," IRE, vol.38, pp. 291-299, March 1950.
2. S. Darlington, "An Introduction to Time-Variable Networks," Proc. of the Midwest Symposium on Circuit Analysis, University of Illinois, pp. 5-1 to 5-25; 1955.
3. S. Darlington, "Nonstationary Smoothing and Prediction Using Network Theory Concepts," IRE Trans. on Circuit Theory, vol. CT-6, Special Supplement, pp. 1-13; May 1959.
4. D.G. Tucker, Circuits with Periodically Varying Parameters, D. Van Nostrand.

5. P. Penfield, "Circuit Theory of Periodically Driven Nonlinear Systems," Proc. of the IEEE, vol. 54, No. 2, pp. 266-280, Feb. 1966.
6. L.E. Franks and I.W. Sandberg, "An Alternative Approach to the Realization of Network Transfer Functions: The N-Path Filter," BSTJ, vol. 39, pp. 1321-1350; September 1960.
7. R.E. Kalman, "Linear Stochastic Filtering Theory--Reappraisal and Outlook," Proc. of the Symposium on System Theory, New York, 1965; pp. 197-205.
8. L.A. Zadeh, "Time-Varying Networks I," Proc. IRE, vol. 49, pp. 1488-1503; October 1961.
9. J.B. Cruz, Jr., "Time-Varying and Active Network Theory," (Part II of "Progress in Circuit Theory--1960-63," Louis Weinberg, editor), IEEE Trans. on Circuit Theory, vol. CT-11, pp. 6-10; March 1964.
10. J.B. Cruz, Jr., and M.E. Van Valkenburg, "The Synthesis of Models for Time-Varying Linear Systems," Proc. of the Symposium on Active Networks and Feedback Systems, Polytechnic Inst. of Brooklyn, pp. 527-544; April 1960.
11. A. Gersho and N. DeClaris, "Duality Concepts in Time-Varying Linear Systems," IEEE Int'l. Convention Record, vol. 12, part 1, pp. 344-356; March 1964.
12. T. Kailath, "Sampling Models for Linear Time-Variant Filters," Mass. Inst. of Technology, RLE Technical Report No. 352; May 1959.
13. E. A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill Book Co., Inc., New York, 1955, Ch. 3.
14. I. W. Sandberg, "Realization of a Class of Periodically Variable Systems," IRE Trans. on Circuit Theory, vol. CT-9, pp. 416-417; December 1962.
15. D.C. Youla, "The Synthesis of Linear Dynamical Systems from Prescribed Weighting Patterns," J. SIAM, vol. 14, pp. 527-549; May 1966.
16. L.M. Silverman and H.E. Meadows, "Equivalence and Synthesis of Time-Variable Linear Systems," to be presented at Fourth Annual Allerton Conference on Circuit and System Theory, University of Illinois, 1966.

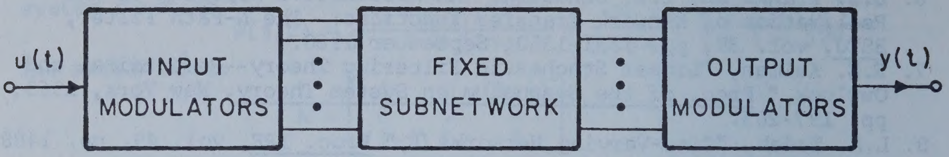

u(t)
INPUT MODULATORS
FIXED SUBNETWORK
OUTPUT MODULATORS
y(t)

FIG. 1 CANONICAL NETWORK

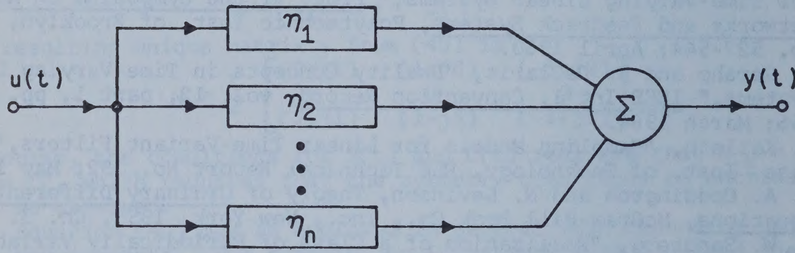

u(t)
η_1
η_2
η_n
Σ
y(t)

FIG. 2 AN N-PATH REALIZATION

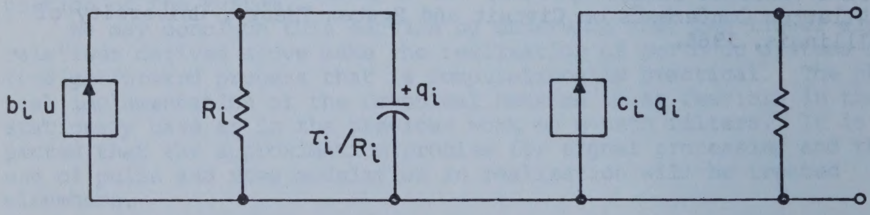

$b_i u$
R_i
$+q_i$
τ_i/R_i
$c_i q_i$

FIG. 3 R C SUBNETWORK

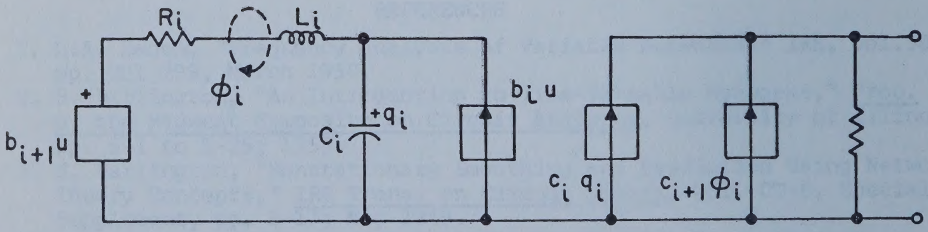

R_i
L_i
ϕ_i
+
$b_{i+1}u$
$+q_i$
C_i
$b_i u$
$c_i q_i$
$c_{i+1}\phi_i$

FIG. 4 RLC SUBNETWORK

STABILITY OF LINEAR TIME-VARYING NETWORKS

K. R. KLECKNER

Department of Engineering, Oakland University, Rochester, Mich.

R. C. WAAG

Rome Air Development Center, Griffiss Air Force Base, Rome, N.Y.

ABSTRACT

Using the second method of Liapunov, both long known and recent stability results are motivated and presented in a logical treatment. First, the necessary and sufficient conditions for stability of the equilibrium for the Λ-matrix form of the network equations are summarized. Then sufficient conditions for network stability are found. This is done in two parts. One outlines conditions that follow directly from the parameter matrices of the various network equation formulations. The other indicates conditions obtained via linear transformations. Stability conditions for fixed networks and two-element-kind time-varying networks follow as special cases of this general treatment.

The results are interpreted in terms of phase space concepts and restrictions on element pumps. Examples of each of the stability conditions are given and comparisons are made between different techniques for determining network stability. In particular, a case is discussed in which the stability of a network is predicted by the present linear transformation technique but is not predicted by the modified energy function method recently given by Kuh.

I. INTRODUCTION

The purpose of this paper is to provide a brief account of the concept of stability as it pertains to three-element-kind time-varying linear networks. Certain scattered results which have been known for some time are presented as special cases of a more general treatment. In addition, certain new results, which have recently been independently obtained by Kuh[1], for the stability of networks directly in terms of the element-parameter matrices are presented. Stability via linear transformations is discussed and shown to be in many cases a more powerful (albeit more difficult computationally) method than that of Kuh. The approach used is Liapunov's second method. Although the general problem of stability of time-varying linear systems is not completely solved because other approaches, based on different definitions of stability which are equally useful for fixed systems, do not yield satisfactory answers in certain special cases for time-varying systems[2,3], nevertheless, the recent revival of interest in the second method of Liapunov has brought the state of the art, from the viewpoint adopted here, to a certain degree of completion.

In the sequel, explicit time dependence will not be indicated except where necessary to avoid confusion, but, unless otherwise stated, strictly positive element values are assumed. The derivative with respect to time of a function f is indicated by $\dot{f}$. Underlined symbols are matrices. The transpose of a matrix $\underline{M}$ is denoted $\underline{M}^T$.

II. BASIC CONCEPTS OF STABILITY IN THE SENSE OF LIAPUNOV

The definition of stability in the sense of Liapunov includes as a special case the common engineering intuitive definition of stability of a system - a stable system being one that, in some sense, returns to its original trajectory after a perturbation. If the original trajectory is a point, the stability becomes stability of the equilibrium of a force-free system. Stability of the equilibrium is the main concern here because it provides information about the solutions composing the fundamental matrix of homogeneous systems. The equilibrium of the system

$$\dot{\underline{Y}} = \underline{A}\,\underline{Y} \tag{1}$$

is said to be stable[4], in the sense of Liapunov if for every $\delta > 0$ there exists a $\eta>0$ and a $t_o \geq 0$ such that every solution for which

$$\underline{Y_o}^T \underline{M}\, \underline{Y_o} \leq \eta$$

at $t = t_o$ remains in the volume

$$\underline{Y}^T(\underline{Y_o},t)\ \underline{M}\ \underline{Y}(\underline{Y_o},t) \leq \delta \text{ for } t > t_o .$$

Necessary and sufficient conditions for the stability of Equation 1 are the existence of two quadratic forms

$$v = \underline{Y}^T\underline{M}\,\underline{Y} \text{ and } w = -\dot{v} = \underline{Y}^T\,\underline{N}\,\underline{Y} \tag{2}$$

where

$$\underline{N} = -(\dot{\underline{M}} + \underline{M}\,\underline{A} + \underline{A}^T\,\underline{M}) \tag{3}$$

which are respectively positive definite and positive semidefinite for every t. If, in addition, for every solution of Equation 1

$$\lim_{t \to \infty} \|\underline{Y}(t)\| = 0$$

the system is said to be asymptotically stable. Necessary and sufficient conditions for asymptotic stability were shown by Gorbunov[5] to be the existence of the positive definite (for every finite, fixed t) forms v and w and the divergence of the integral

$$\int_{t_0}^{t} \lambda_0(\xi)\, d\xi$$

as $t \to \infty$, where λ_0 is the smallest eigenvalue of $\underline{N}\,\underline{M}^{-1}$. These necessary and sufficient conditions may be equivalently stated in terms of the existence of a positive-definite decrescent* quadratic form w of Equation 2, for every t. The quadratic form v is commonly known as a Liapunov function. Asymptotic stability of the homogeneous system can also be determined from bounds on the system coefficient matrix and on the impulse response[6].

The Liapunov function v and its rate of change are encountered in the Lagrangian and Hamiltonian formulations of the equations of linear networks, where they are associated with energy and power. In phase-space analysis, they have the interpretation of distance and rate of closure. Thus, asymptotic stability implies the existence of a special coordinate system in which the distance to the origin is decreasing or a set of variables such that the total energy of the system is decreasing.

The practical application of Liapunov's second method involves choosing matrix $\underline{M}$, computing the resulting matrix $\underline{N}$, and then determining the ranges of system parameters for which $\underline{N}$ is positive semidefinite or definite. The ranges so determined depend upon the initial choice of $\underline{M}$, and one seeks $\underline{M}$'s which yield the largest stability domains. The following sections consider four simple choices for $\underline{M}$ when analyzing time-varying linear networks. Which of them will be most advantageous depends upon both the network element time variations and the method of formulation of the equilibrium equations.

---

* i.e., $\lim_{\|Y\| \to 0} v(\underline{Y},t) = 0$ holds uniformly in t.

III. SUFFICIENT STABILITY CONDITIONS DIRECTLY FROM PARAMETER MATRICES

The homogeneous network equations of two- or three-element-kind networks can always be written in the form

$$\underline{A}_1\dot{\underline{Y}} + \underline{A}_o\underline{Y} = \underline{0} \tag{4}$$

or the form

$$\frac{d}{dt}(\underline{\hat{A}}_1\,\underline{Y}) + \underline{\hat{A}}_o Y = \underline{0} \quad , \tag{5}$$

where the specific form of the $\underline{A}$'s depends on the basic method of formulation of the equations. When put into the canonical form of Equation 1, these expressions become, respectively,

$$\dot{\underline{Y}} = -\underline{A}_1^{-1}\underline{A}_o\underline{Y} \tag{6}$$

and

$$\dot{\underline{Y}} = -\hat{A}_1^{-1}(\hat{A}_o + \hat{A}_1)\underline{Y} \tag{7}$$

The equations for a three-element-kind network established by the **Guillemin**[7] method are cases where $\underline{A}_1$ and $\underline{\hat{A}}_1$ are symmetric. In many practical cases, $\underline{A}_1$ and $\underline{\hat{A}}_1$ will also be positive definite (e.g., networks for which all the eigenvalues of $\underline{A}_1$ and $\underline{\hat{A}}_1$ are greater than a fixed bound $\delta > 0$ that is independent of t). In such cases, for the system of Equations 4 and 6, choose

$$\underline{M} = \frac{1}{2}\underline{A}_1 \tag{8}$$

and, from Equation 3, obtain

$$\underline{N} = \frac{1}{2}(\underline{A}_o + \underline{A}_o^T) - \frac{1}{2}\underline{\dot{A}}_1 \tag{9}$$

For the system of Equations 5 and 7, choose

$$\underline{M} = \frac{1}{2}\underline{\hat{A}}_1 \tag{10}$$

which then yields

$$\underline{N} = \frac{1}{2}(\underline{\hat{A}}_o + \underline{\hat{A}}_o^T) + \frac{1}{2}\underline{\dot{\hat{A}}}_1 \tag{11}$$

Thus, in the cases where $\underline{N}$ is positive semidefinite, the network is stable. If $\underline{A}_1$ and $\underline{\hat{A}}_1$ are also decrescent (e.g., if all their eigenvalues are bounded above by $\eta > 0$), then asymptotic stability can be inferred from the positive definiteness of $\underline{N}$.

If the network equations are obtained from the Hamiltonian formulation[8], leading to Equation 4, $\underline{A}_o$ will be symmetric. If it is also positive definite, choose

$$\underline{M} = \frac{1}{2}A_o \tag{12}$$

and obtain

$$\underline{N} = \frac{1}{2}\,\underline{A}_o\,(\underline{A}_1^{-1T} + \underline{A}_1^{-1})\,\underline{A}_o - \frac{1}{2}\,\dot{\underline{A}}_o \quad .$$

Stability of the network is then guaranteed by the positive semidefiniteness of the congruently related matrix

$$\underline{N}^* = \underline{A}_o^{-1}\,\underline{N}\,\underline{A}_o^{-1} = \frac{1}{2}\,(\underline{A}_1^{-1T} + \underline{A}_1^{-1}) - \frac{1}{2}\,\underline{A}_o^{-1}\,\dot{\underline{A}}_1\,\underline{A}_o^{-1}$$

$$= \frac{1}{2}\,(\underline{A}_1^{-1T} + \underline{A}_1^{-1}) + \frac{1}{2}\,\frac{d}{dt}\,(\underline{A}_o^{-1}) \quad . \qquad (13)$$

Again, the additional requirements of the decrescence of $\underline{A}_o$ and the positive definiteness of $\underline{N}^*$ will guarantee asymptotic stability.

As an example of the type of sufficient stability conditions obtained for Equations 5 and 7 from Equations 10 and 11, consider the reduced equations of an RLC network written in terms of node voltages $\underline{V}$, loop currents $\underline{J}$, the loop-parameter matrices $\underline{R}$ and $\underline{L}$, and the node-parameter matrices $\underline{C}$ and $\underline{G}$, i.e.,

$$\frac{d}{dt}\left\{\left[\begin{array}{c:c}\underline{L} & \underline{0}\\ \hdashline \underline{0} & \underline{C}\end{array}\right]\left[\begin{array}{c}\underline{J}\\ \hdashline \underline{V}\end{array}\right]\right\} + \left[\begin{array}{c:c}\underline{R} & \underline{\gamma}\\ \hdashline -\underline{\gamma}^T & \underline{G}\end{array}\right]\left[\begin{array}{c}\underline{J}\\ \hdashline \underline{V}\end{array}\right] = \underline{0}$$

Then, if $\underline{L}$ and $\underline{C}$ are positive definite, one has

$$\underline{M} = \frac{1}{2}\left[\begin{array}{c:c}\underline{L} & \underline{0}\\ \hdashline \underline{0} & \underline{C}\end{array}\right]$$

and

$$\underline{N} = \left[\begin{array}{c:c}\underline{R} & \underline{0}\\ \hdashline \underline{0} & \underline{G}\end{array}\right] + \frac{1}{2}\left[\begin{array}{c:c}\dot{\underline{L}} & \underline{0}\\ \hdashline \underline{0} & \dot{\underline{C}}\end{array}\right]$$

With $\underline{X}^T = \left[\underline{J}^T \;\vdots\; \underline{V}^T\right]$, the quadratic form $v = \underline{X}^T\,\underline{M}\,\underline{X}$ represents the instantaneous stored field energy, and the first term of the quadratic form $w = \underline{X}^T\,\underline{N}\,\underline{X}$ represents the power dissipated. The second term of w is the rate of energy removal by the reactive-element pumps. The Liapunov equation $\dot{v} + w = 0$ thus represents conservation of energy, and the sufficient stability condition that $\underline{N}$ be positive semidefinite is a condition, based on network resistances and topology, that the rate of energy addition via the pumps not be too large.

As a second example, consider an RLC network for which the reduced equations can be put into the form

$$\left[\begin{array}{c:c}\underline{R} & \underline{0}\\ \hdashline \underline{0} & \underline{G}\end{array}\right]\left[\begin{array}{c}\underline{Q}\\ \hdashline \dot{\underline{\phi}}\end{array}\right] + \left[\begin{array}{c:c}\underline{S} & \underline{\kappa}\\ \hdashline \underline{\mu} & \underline{\Gamma}\end{array}\right]\left[\begin{array}{c}\underline{Q}\\ \hdashline \underline{\phi}\end{array}\right] = \underline{0}$$

where $\dot{\underline{Q}} = \underline{J}$, $\dot{\underline{\Phi}} = \underline{V}$, $\underline{S}$ and $\underline{R}$ are loop elastance and resistance matrices, and $\underline{G}$ and $\underline{\Gamma}$ are node conductance and reciprocal inductance matrices. Then, if $\underline{R}$ and $\underline{G}$ are positive definite, use Equations 8 and 9 to get

$$\underline{M} = \frac{1}{2} \left[\begin{array}{c|c} \underline{R} & \underline{0} \\ \hline \underline{0} & \underline{G} \end{array}\right]$$

and

$$\underline{N} = \left[\begin{array}{c|c} \underline{S} & \underline{r} \\ \hline \underline{r}^T & \underline{\Gamma} \end{array}\right] - \frac{1}{2} \left[\begin{array}{c|c} \dot{\underline{R}} & \underline{0} \\ \hline \underline{0} & \dot{\underline{G}} \end{array}\right], \qquad r = \frac{1}{2}(\kappa + \mu^T)$$

In this case, $\underline{M}$ is associated with the dissipation. It may be more convenient, however, to associate $\underline{M}$ with a generalized distance in phase space since the former choice does not provide a clear energy balance interpretation for the equation $\dot{v} + w = 0$, as it does in the previous formulation. Nonetheless, the sufficient stability condition of $\underline{N}$ positive semidefinite may be considered to be a condition, based on the network storage elements and topology, on the resistance pumps that guarantees network stability.

Similar results for two-element-kind networks have been obtained by Darlington[9]. Under the assumption of periodicity, Darlington has also shown how expressions like these lead to bounds on the real parts of the characteristic exponents. Desoer and Paige[10] have given equivalent conditions for lossy two-element-kind networks for the more restrictive condition that only one of the parameter matrices varies. For these special cases, stability is most easily demonstrated by the foregoing procedure.

IV. SUFFICIENT STABILITY CONDITIONS VIA LINEAR TRANSFORMATIONS

As pointed out by Kuh[1], the preceding stability conditions directly in terms of parameter matrices are easily applied but not particularly sharp. The exchange between sharpness and ease of calculation is characteristic of Liapunov's second method. From the standpoint of simplicity, the best choice for $\underline{M}$ is a constant diagonal matrix. Then, from Equation 3,

$$\underline{N} = -\frac{1}{2}(\underline{A} + \underline{A}^T) \text{ with } \underline{M} = \frac{1}{2}\underline{I}, \tag{14}$$

($\underline{I}$ is a unit matrix of appropriate dimension) and one need only find the eigenvalues of this relatively simple matrix to obtain sufficient conditions for stability. Unfortunately, this simple choice for $\underline{M}$ does not usually yield sharp stability bounds.

A more general formulation of the stability problem is obtained by incorporating an independently specifiable time function in the Liapunov function. Kuh chooses $v = f(t)\,E(t)$ where $E(t)$ is what would be the total stored energy if the network were time-invariant, and $f(t)$ is a positive function chosen in a special way. With this choice he is able to improve upon the bounds obtained via the parameter matrices.

Another approach is to consider linear transformations of the original coordinate system of the network variables[11]. In this case, set

$$\underline{M} = \frac{1}{2}\,\underline{P}^T\underline{P}$$

where $\underline{P}$ is real and nonsingular. If it is now assumed that $\underline{A}$ has distinct eigenvalues (there is no real loss of generality here, since the case in which $\underline{A}$ has repeated eigenvalues can be regarded as the limit of distinct eigenvalues being shifted so as to overlap), then $\underline{P}$ can be chosen as the matrix which transforms $\underline{A}$ into the matrix.

$$\underline{Z} = \begin{bmatrix} \rho_1 & & & & & & \\ & \rho_2 & & & & \underline{0} & \\ & & \ddots & & & & \\ & & & \rho_k & & & \\ & & & & \begin{bmatrix} \sigma_{k+1} & \omega_{k+1} \\ -\omega_{k+1} & \sigma_{k+1} \end{bmatrix} & & \\ & \underline{0} & & & & \ddots & \\ & & & & & & \begin{bmatrix} \sigma_m & \omega_m \\ -\omega_m & \sigma_m \end{bmatrix} \end{bmatrix}$$

where the ρ_i are the real eigenvalues of $\underline{A}$ and the $\sigma_i \pm j\omega_i$ are the complex eigenvalues of $\underline{A}$, i.e.,

$$\underline{A} = \underline{P}^{-1}\,\underline{Z}\,\underline{P}$$

Equation 3 then gives for $\underline{N}$:

$$\underline{N} = -\frac{1}{2}\left[\underline{P}^T(\underline{Z} + \underline{Z}^T)\,\underline{P} + \frac{d}{dt}(\underline{P}^T\,\underline{P})\right]$$

Using the fact that $\underline{Z} + \underline{Z}^T = 2\,\mathrm{Re}\left\{\underline{\lambda}\right\}$ where $\underline{\lambda}$ is the diagonal matrix of eigenvalues of $\underline{A}$ gives:

$$\underline{N} = -\left[\underline{P}^T\,\mathrm{Re}\left\{\underline{\lambda}\right\}\underline{P} + \frac{d}{dt}\,(\underline{P}^T\,\underline{P})\right] \qquad (15)$$

In the general case $\dot{\underline{P}} \neq \underline{0}$, and hence $\underline{P}$ does not diagonalize $\underline{N}$. Among the special cases in which $\underline{P}$ does vanish are fixed-parameter RLC networks and time-variable two-element-kind networks in which all elements of one kind are fixed[10]. In such cases the real parts of the eigenvalues of $\underline{A}$ being nonpositive guarantees stability. The still more special case where $\underline{A}$ is symmetric results in $\underline{P}^T = \underline{P}^{-1}$, i.e., $\underline{M} = \frac{1}{2}\,\underline{I}$, and Equations 14 and 15 clearly yield identical stability conditions.

For fixed-parameter networks, the parameter matrix method, the method of Kuh (with $f(t) \equiv 1$), and the preceding linear transformation method are essentially equivalent. The choice of $\underline{M} = \frac{1}{2}\,\underline{I}$ is still generally inferior, however, unless $\underline{A}$ is symmetric. For example, if

$$A = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -2 & 2 \\ 1 & -2 & -1 \end{bmatrix}$$

then $\underline{N}$ defined by Equation 15 is positive definite whereas $\underline{N}$ defined by Equation 14 is not. In fact, if λ is an eigenvalue of $\underline{A}$ and $\beta_1, \beta_2, \ldots \beta_n$ are the eigenvalues of $\frac{1}{2}\,(\underline{A} + \underline{A}^T)$ and are so arranged that

$$\beta_1 \ge \beta_2 \ge \cdots \ge \beta_n,$$

a theorem of Bendixson[12] stated that

$$\beta_1 \ge \mathrm{Re}\left\{\underline{\lambda}\right\} \ge \beta_n \qquad (16)$$

This shows that Equation 15 can give an improvement over Equation 14 with respect to sufficient stability conditions.

For time-variable networks, Equation 15 can again yield sharper stability conditions than Equation 14, although it is now unclear whether it always does. For example, the second-order $\underline{A}$ matrix

$$\underline{A} = -\alpha\underline{I} + 8\begin{bmatrix} (\cos t + \sin t)(1 - \frac{1}{2}\sin 2t) & (\cos t - \sin t)\sin 2t \\ \frac{1}{4}(\cos t - \sin t)\sin 2t & \frac{1}{2}(\cos t + \sin t)\sin 2t \end{bmatrix}$$

where $\underline{I}$ is a 2 x 2 unit matrix. Choosing $\underline{M} = \frac{1}{2}\,\underline{I}$ leads one to investigate the definiteness of the matrix

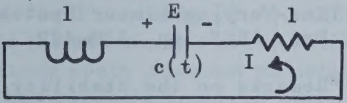
$$\underline{N}_1 = \alpha\underline{I} + \begin{bmatrix} 8(\cos t + \sin t)(1 - \frac{1}{2}\sin 2t) & 5(\cos t - \sin t)\sin 2t \\ 5(\cos t - \sin t)\sin 2t & 4(\cos t + \sin t)\sin 2t \end{bmatrix}$$

Choosing $\underline{M} = \frac{1}{2}\,\underline{P}^T\underline{P}$ with $\underline{P} = \begin{bmatrix} \cos t & 2\sin t \\ \frac{1}{2}\sin t & \cos t \end{bmatrix}$ leads to the matrix

$$\underline{N}_2 = \alpha\underline{I} + \begin{bmatrix} 8\cos t & 3/4 \\ 3/4 & 8\sin t \end{bmatrix}$$

For $\alpha > 10$, both $\underline{N}_1$ and $\underline{N}_2$ predict stability; for $8.07 < \alpha < 10$, $\underline{N}_2$ predicts stability, but $\underline{N}_1$ does not.

Because of the manners in which the network equations and Liapunov functions are formed, the parameter-matrix method and the modified energy function method both require that all inductances and capacitances be strictly positive. The linear transformation method is not so restricted. As an example, consider the following network, in which the



$$c(t) = \frac{16}{9}e^{-at} \quad (0 < a < \tfrac{1}{2})$$

capacitance variation precludes use of the first two approaches. The network equations are:

$$\begin{bmatrix}\dot{E}\\ \dot{I}\end{bmatrix} = \begin{bmatrix} a & -\frac{9}{16}e^{at} \\ 1 & -1\end{bmatrix}\begin{bmatrix}E\\ I\end{bmatrix}.$$

The matrix $\underline{P}$ is readily computed to be

$$\underline{P} = \begin{bmatrix} 1 & -\frac{1}{2}(a+1) - \frac{1}{2}w \\ -1 & \frac{1}{2}(a+1) - \frac{1}{2}w \end{bmatrix}$$

where $w = +\sqrt{\frac{9}{4}e^{at} - (a+1)^2}$, and $\mathrm{Re}\{\lambda\}$ is given by

$$\mathrm{Re}\{\lambda\} = \frac{1}{2}(a-1)\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}.$$

Equation 15 then yields

$$\underline{N} = \begin{bmatrix} \frac{1}{2}(1-a) & \frac{1}{4}(a^2-1) \\ \frac{1}{4}(a^2-1) & \frac{9}{32}(1-3a)e^{at} \end{bmatrix}.$$

For $a < 0.165$, $\underline{N}$ is positive definite, indicating stability. (Note that asymptotic stability cannot be inferred as $\underline{M} = \frac{1}{2}\underline{P}^T\underline{P}$ is not decrescent.)

V. CONCLUSION

The prediction of stability of time-varying RLC networks by means of Liapunov's second method becomes increasingly sharper as one is willing to undertake increasingly tedious calculations. The simplest, and weakest, approach is to choose $\underline{M}$ to be a constant diagonal matrix. If no conclusions can be drawn, try the parameter matrix method or the modified energy function method of Kuh (in that order) if they are applicable. If matters are still undecided, the linear transformation approach may clear them up.

REFERENCES

1. E. S. Kuh, "Stability of Linear Time-Varying Networks-The State-Space Approach," *IEEE Trans. on Circuit Theory*, *CT-12*, June 1965, pp. 150-157.

2. R. E. Kalman, "On the Stability of Time-Varying Linear Systems," *IRE Trans. on Circuit Theory*, *CT-9*, Dec. 1962, pp. 420-422.

3. T. F. Bridgeland and R. E. Kalman, "Remarks on the Stability of Linear Systems," *IEEE Transactions on Circuit Theory*, CT-10, Dec. 1963, pp. 539-542.

4. V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Translated by S. Lefshetz, Princeton University Press, Princeton, New Jersey, 1960, Chapter 3.

5. A. D. Gorbunov, "On Conditions for Asymptotic Stability of the Zero-Solution of a System of Ordinary Linear Differential Equations," *Vestnik Moskovskogo Universitia, Serijo Fizikomate Maticeskich i estestvennich Nauk*., 8, 1953, pp. 46-55. (See also: W. Hahn, *Theory and Application of Liaponov's Second Method*, Translated by S. H. Lehnigk & H. H. Hosenthien, Prentice-Hall, Englewood Cliffs, N. J. 1963, Chapter 5.)

6. R. E. Kalman and J. E. Bertram, "Control System Analysis and Design Via the "Second Method of Liaponov", *J. Basic Engineering (Trans. ASME)*, *820*, June 1960, pp. 371-393.

7. E. A. Guillemin, "The Normal Coordinate Transformation of a Linear System With an Arbitrary Loss Function," *J. Math*, *Phys.*, v39, n2, July 1960.

8. R. C. Waag, "Investigation of Linear Time-Varying Networks Using Classical Methods," PhD. Thesis, Cornell University, Ithaca, N. Y., 1965.

9. S. Darlington, "Linear Time-Varying Circuits--Matrix Manipulation, Power Relations, and Some Bounds on Stability," *Bell System Tech. Journal*, 43, Nov. 1963, pp. 2575-2608.

10. C. A. Desoer and A. Paige, "Linear Time-Varying G-C Networks: Stable and Unstable," *IEEE Trans. on Circuit Theory*, *CT-10*, June 1963, pp. 180-190.

11. R. A. Rohrer, "Transformation Techniques for the Application of Liapunov's Direct Method to Linear Systems," *IEEE Trans. on Circuit Theory*, *CT-11*, Mar 64, pp. 171-173.

12. I. Bendixson, "Sur les Racines d'une Equation Fondamentale," *Acta Mathematica*, *25*, 1902, pp. 359-365.

CONSTANT RESISTANCE, WIDE-SENSE SOLVABILITY, AND SELF-DUALITY[*]

C. A. Desoer,[†] R. W. Newcomb,[§] and K. K. Wong[†]

Using his concept of system function, Zadeh has shown that every self-dual one-port made of linear time-varying elements is a constant resistance one-port [1]. Recently we gave instances of constant resistance one-ports that have nonlinear, time-varying elements. Some of these one-ports are self-dual networks [2-4]. Here we give a precise condition for the truth of the statement "every self-dual one-port is constant resistance and conversely every constant resistance one-port is self-dual." This proposition has recently acquired more importance since wide classes of self-dual one-ports can easily be generated [4-5]. This paper is an extension of a previous paper [6] in that we adopt exclusively a black-box point of view, and it proves the equivalence completely.

We assume throughout that all one-ports under consideration have been created at $t = -\infty$ and that at the time of their creation they are in their zero-state. Similarly, any interconnection of one-ports is assumed to be done at $-\infty$. As a consequence, *all* waveforms under consideration are defined on $(-\infty, \infty)$.

By definition, a one-port $\mathcal{N}$ is specified as the set of all voltage current pairs $[v(\cdot), i(\cdot)]$ it allows. A one-port $\mathcal{N}^*$ is said to be the *dual* of $\mathcal{N}$ whenever the following condition holds: $[f, g] \in \mathcal{N}^*$ if and only if $[g, f] \in \mathcal{N}$. A one-port $\mathcal{N}$ is said to be *self-dual* whenever $[f, g] \in \mathcal{N}$ implies $[g, f] \in \mathcal{N}$. This point of view amounts to thinking of a one-port as a binary relation on some function space [7, p. 9]; the converse relation is the dual one-port; a one-port is self-dual if and only if its defining relation is symmetric. Given a one-port $\mathcal{N}$, we define the *augmented one-port* $\mathcal{N}_a$ by its ordered pairs: $[v + i, i] \in \mathcal{N}_a$ when and only when $[v, i] \in \mathcal{N}$; $\mathcal{N}_a$ has an obvious interpretation given in Fig. 1a. Any voltage $e(\cdot)$ such that $e = v + i$ for some $[v, i] \in \mathcal{N}$ is called an *allowed* voltage of $\mathcal{N}_a$. We now slightly extend the concept of solvability

---

[*] The research reported herein was supported in part by the National Science Foundation under Grant GK-716, in part by the U. S. Army Research Office--Durham under Grant DA-ARO-D-31-124 (both grants to the Electronics Research Laboratory, University of California, Berkeley) and in part by the Air Force Office of Scientific Research under Grant AF-AFOSR-337-63 to Stanford University, Stanford, California.

[†] Department of Electrical Engineering and Electronics Research Laboratory, College of Engineering, University of California, Berkeley.

[§] Department of Electrical Engineering, Stanford University, Stanford, California.

[8, p.113; 9, p.9] by considering only a restricted class of $e(\cdot)$'s, namely those allowed by $\mathcal{N}_a$. $\mathcal{N}$ is said to be wide-sense solvable (abbreviated as w.s. solvable) if for all allowed $e(\cdot)$, the equation $i(\cdot) + v(\cdot) = e(\cdot)$ has a unique solution $[v(\cdot), i(\cdot)] \in \mathcal{N}$. Physically, w.s. solvability means that if a voltage source (whose voltage $e(\cdot)$ is an allowed voltage of $\mathcal{N}_a$) is connected to $\mathcal{N}_a$, then the port voltage and port current of $\mathcal{N}$ are uniquely determined. Note that the nullator is not solvable in the sense of Youla et al. [8] and Newcomb [9] but is w.s. solvable.

When we consider the one-port $\mathcal{N}$ as a "constant resistance one-port" we only allow $\mathcal{N}$ to be connected to one-ports $\mathcal{N}'$ such that the connection $\mathcal{N} - \mathcal{N}'$ is determinate, i.e., the port voltage $v(\cdot)$ and the port current $i(\cdot)$ of $\mathcal{N}$ are uniquely determined. Such one-ports $\mathcal{N}'$ are said to be compatible with $\mathcal{N}$. If $\mathcal{N}$ is w.s. solvable, then the series connection of a one-ohm resistor and a voltage source e where e is an allowed voltage of $\mathcal{N}_a$ is a one-port compatible with $\mathcal{N}$. If all connections $\mathcal{N} - \mathcal{N}'$ where $\mathcal{N}'$ is compatible with $\mathcal{N}$ have the property that the port voltage $v(\cdot)$ (of $\mathcal{N}$) is equal to the port current $i(\cdot)$ (of $\mathcal{N}$), we say that $\mathcal{N}$ is constant resistance. By including a scale factor, this definition can be extended to include the case where for all such connections, $v(\cdot) = ki(\cdot)$, where k is a fixed non-zero real number independent of $i(\cdot)$, $v(\cdot)$, and t. We want now to prove the

Theorem. A one-port $\mathcal{N}$ is constant resistance if and only if $\mathcal{N}$ is w.s. solvable and self-dual.

Proof. 1. Wide-sense solvability and self-duality imply constant resistance. Let $\mathcal{K}(v)$ denote any member of $\{\tilde{i} : [v, \tilde{i}] \in \mathcal{N}\}$; $\mathcal{K}$ is not necessarily a function but describes the relation defining $\mathcal{N}$. From Fig. 1b, and the w.s. solvability assumption, the equation

$$e = v + \mathcal{K}(v) \tag{1}$$

has a unique solution for all allowed e. Figure 1c shows the dual of Fig. 1b; then, with the notations shown in Fig. 1c, $\hat{i} = v$ and $\hat{v} = i$, by duality. By self-duality, $i = \mathcal{K}(v)$ implies $v = \mathcal{K}(i)$, or what is the same $\hat{i} = \mathcal{K}(\hat{v})$. From Fig. 1c, KCL gives $\hat{j} = e = \hat{v} + \hat{i}$,

hence
$$e = \hat{v} + \mathcal{K}(\hat{v}) . \tag{2}$$

Since for all allowed e, this equation has a unique solution, Eqs. (1) and (2) imply that $v = \hat{v}$. Hence, $v = i$ and the one-port $\mathcal{N}$ is equivalent to a one-ohm resistor when it is driven by any allowed voltage source in series with a one-ohm resistor. That it is equivalent to a one-ohm resistor under all compatible connections is obvious by contradiction: suppose it were not true, then there would exist a compatible one-port $\mathcal{N}'$ such that the connection $\mathcal{N} - \mathcal{N}'$ has a solution $[\tilde{v}, \tilde{i}]$ with $\tilde{v} \neq \tilde{i}$. Now consider $\mathcal{N}_a$ driven by the allowed voltage source $\tilde{e} \triangleq \tilde{v} + \tilde{i}$: by the w.s. solvability assumption and the definition of $\tilde{v}$, $\tilde{i}$ there is only one possible port voltage and port current, namely, $\tilde{v}$ and $\tilde{i}$. But the previous proof requires $\tilde{i} = \tilde{v}$. This is a contradiction, hence $\mathcal{N}$ is equivalent to a one-ohm resistor under all compatible connections, i.e., $\mathcal{N}$ is constant resistance.

2. Constant resistance implies self-duality and w. s. solvability. Let $[v_0, i_0]$ be an arbitrary pair of $\mathcal{N}$. Consider the one-port $\mathcal{N}'_0$ shown in Fig. 2: the current source i_0 and the voltage source v_0 of $\mathcal{N}'_0$ are independent sources; the nullator admits only the pair $[0, 0]$. By KCL, KVL and the defining relations of the elements of $\mathcal{N}'_0$, the one-port $\mathcal{N}'_0$ admits only one pair $[v_0, -i_0]$. The connection $\mathcal{N} - \mathcal{N}'_0$ has a unique solution: $[v_0, i_0]$, i.e., $\mathcal{N}'_0$ is compatible with $\mathcal{N}$. By the constant resistance assumption, $v_0 = i_0$. Thus we have shown that, for all $[v, i] \in \mathcal{N}$, $v = i$. This implies that $\mathcal{N}$ is self-dual. Given any allowed voltage e, the only solution of $e = v + i$, with $[v, i] \in \mathcal{N}$, is $v = i = (e/2)$, i.e., $\mathcal{N}$ is w. s. solvable.

Q. E. D.

It follows from the proof of the theorem that $\mathcal{N}$ is constant resistance if and only if $v = i$ for all $[v, i] \in \mathcal{N}$.

Remarks.

a. By interpreting all voltages and all currents as n-vectors one sees that all definitions and derivations are still valid, consequently the theorem holds for n-ports.

b. It should be stressed that the point of view adopted in this paper is strictly black box: only the port voltage and the port current are observable and the set of all pairs $[v, i]$ constitute the complete description of the one-port. An immediate consequence is that the theorem applies to any one-port: its elements may be lumped or distributed, active or passive, linear or nonlinear, time-varying or time-invariant. On the other hand one should keep in mind that the black box self-duality defined here does not imply, for example, that the graph of the network inside the box is a self-dual graph. For example, the linear time-invariant network of Fig. 3 of a previous paper [3] is self dual in the present (black box) sense but its graph is not a self-dual graph.

c. Given an arbitrary one-port $\mathcal{N}$ and its dual $\mathcal{N}^*$ (as defined in this paper), it is possible to use $\mathcal{N}$ and $\mathcal{N}^*$ as elements to obtain constant resistance one-ports. (See Examples 1 and 2 of Sec. III in Ref. [4].)

d. Let a be a fixed real number. If in the one-port shown in Fig. 2 we set $v_0(t) = -i_0(t) = a$ for all t, we then obtain a constant resistance one-port: indeed, its only pair is $[a, a]$. With $a = 0$, we see that the nullator is a constant resistance one-port.

e. The following one-port $\mathcal{N}_1$ shows that self-duality implies neither constant resistance nor w. s. solvability. Let $\mathcal{N}_1$ admit only constant voltages and currents and let its admissible pairs be $[V, I]$ where either $V = 2I$ or $V = 2^{-1}I$. $\mathcal{N}_1$ is clearly self-dual but neither constant resistance nor w. s. solvable.

References

[1] L. A. Zadeh, "Constant-resistance networks of the linear varying-parameter type," Proc. IRE, Vol 39, pp. 688-691, June 1951.

[2] C. A. Desoer and K. K. Wong, "Constant resistance nonlinear time varying one-ports," Proc. IEEE, 53, 11, pp. 1744-1745, Nov. 1965.

[3] C. A. Desoer and K. K. Wong, "Constant resistance one-ports with nonlinear time-varying elements," IEEE International Convention Record, pp. 233-234, March 1966.

[4] C. A. Desoer and K. K. Wong, "Constant resistance one-ports which include nonlinear time-varying elements," IEEE Transactions on Circuit Theory, to appear.

[5] P. M. Lin, School of Electrical Engineering, Purdue University. Private communication.

[6] C. A. Desoer and K. K. Wong, "Self-duality and constant resistance one-ports," Letter to the Editor, Proc. IEEE, to appear.

[7] M. A. Harrison, Introduction to Switching and Automata Theory, McGraw-Hill, New York, 1965.

[8] D. C. Youla, L. J. Castriota, and H. J. Carlin, "Bounded real scattering matrices and the foundations of linear passive network theory," IRE Transactions on Circuit Theory, Vol. CT-6, No. 1, pp. 102-124, March 1959.

[9] R. W. Newcomb, "The foundations of network theory," Transactions of the Institution of Engineers, Australia, Vol. EM 6, No. 1, pp. 7-12, May 1964.

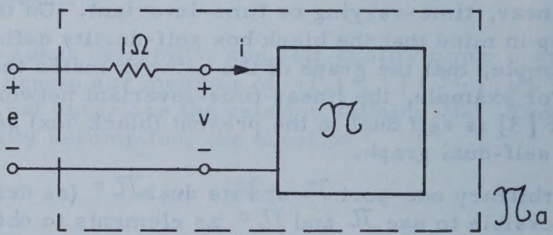


Fig. 1a. Physical relation between e, i, and v.

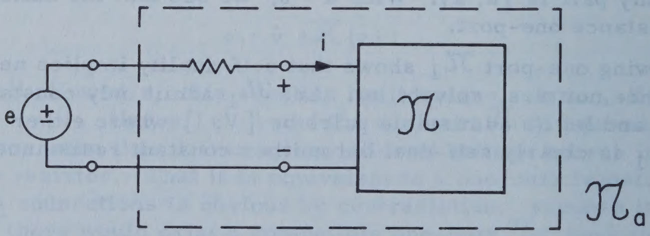


Fig. 1b. Circuit required for testing solvability.

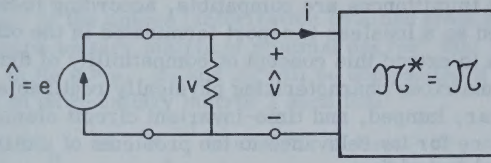

i
+
$\hat{j} = e$
1v
$\hat{v}$
−
$\mathcal{N}^* = \mathcal{N}$

Fig. 1c. The dual of (b). $\mathcal{N}^*$, the dual of $\mathcal{N}$, is identical to $\mathcal{N}$ since $\mathcal{N}$ is self dual.

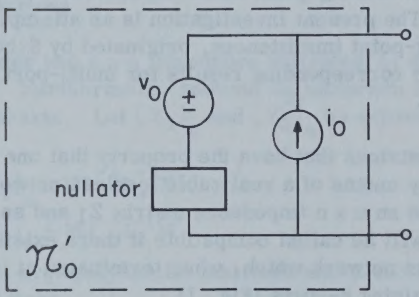

v_0
±
i_0
nullator
$\mathcal{N}_0'$

Fig. 2. $\mathcal{N}_0'$ is compatible with $\mathcal{N}$, and $\mathcal{N}_0'$ has only one admissible voltage current pair, namely $[v_0, -i_0]$.

COMPATIBILITY OF IMMITTANCE MATRICES*

S. K. MULLICK
IBM Components Division, East Fishkill Facility, Hopewell Junction, N.Y.

ABSTRACT

Two driving-point immittances are compatible, according to Schoeffler, if one of them can be realized as a lossless two-port terminated in the other. The purpose of the present study is to extend this concept of compatibility of driving-point immittances to immittance matrices characterizing physically realizable networks composed of passive, linear, lumped, and time-invariant circuit elements. The study has practical importance for its relevance to the problems of multi-port matching and lossless reciprocal imbedding.

INTRODUCTION

The algebra of matrices of rational functions is a natural tool for the study of multi-port networks. Matrix methods play a major role in extending the concepts and results pertaining to driving-point immittances of two terminal networks to the multi-terminal case. The present investigation is an attempt to extend the notion of compatibility of driving-point immittances, originated by Schoeffler,[1] to immittance matrices, and to derive corresponding results for multi-port structures characterized by these matrices.

Two immittance matrices that have the property that one of them can be transformed into the other by means of a realizable lossless network are called compatible. Specifically, given an n x n impedance matrix Z_1 and an m x m impedance matrix Z_2, Z_1 and Z_2 will be called compatible if there exists a (m + n)-port physically realizable lossless network which, when terminated at its m-ports by $Z_2(s)$, yields $Z_1(s)$ at its remaining n-ports (Fig. 1).

TWO THEOREMS ON COMPATIBILITY

The compatibility conditions can be derived in two ways. In the first method, a scattering matrix normalized to the multi-port load network is defined,[2,3] and an existence theorem for such a scattering matrix is established. Restricting this matrix to be para-unitary, then, yields the desired conditions. This approach is essentially a generalization of that of Wohler's[4] for the study of compatibility of driving-point immittances. The second method is valid when the immittance matrices are of the same order. This procedure makes use of the transmission matrix representation of the coupling structure as an intermediate step and follows similar lines of argument to those laid down by Bayard[5] for synthesis of n-ports as lossless structures terminated in appropriate number of resistances.

---

*This paper is based on a thesis which the author submitted to the Graduate School of Cornell University, Ithaca, N.Y., in partial fulfillment of the requirements for the Ph.D. degree, June 1965. This work was supported by the Air Force Systems Command, Wright-Patterson Air Force Base, Dayton, Ohio.

The results obtained by these approaches can be summarized in the following two theorems:

Theorem 1: Consider two pr impedance matrices Z_1 and Z_2 of orders n and m ($n \geq m$) respectively, of which Z_2 is a non-Foster matrix. Let their even parts, $[Z_1]_e$ and $[Z_2]_e$ respectively, to be factored as follows:

(a) $[Z_1]_e = K_*'K$ is the spectral factorization[6,7]

(b) $[Z_2]_e = H_2H_{2*}'$ is the unique factorization obtained from spectral factorization with the help of a para-unitary matrix of minimal degree[8] that renders the matrix $H_2H_{2*}'^{-1}$ analytic in $\mathrm{Re}\, s > o$. Z_1 and Z_2 will be compatible if and only if there exists a regular m x m para-unitary matrix Υ such that

1. $Y_{12} = [Z_1 + 1_n]^{-1}K'\Upsilon'H_2^{-1}$ is analytic in $\mathrm{Re}\, s > o$;
2. $Y_{22} = 1/2\, H_2^{-1}[H_2H_2'^{-1} - \Upsilon K'^{-1}(1_n - Z_1)(1_n + Z_1)^{-1}K'\Upsilon']H_2^{-1}$ is analytic in $\mathrm{Re}\, s > o$;
3. $[1_m - Z_2]Y_{12}'$ is analytic on the $j\omega$ axis; and
4. $[Z_2 + 1_m][1_m - (Z_2 - 1_m)Y_{22}]$ has simple poles on the $j\omega$ axis with positive semidefinite residue matrices.

Theorem II: Consider two n x n impedance matrices Z_1 and Z_2 whose even parts are non-singular. Furthermore, assume Z_2 minimum reactive and non-singular on the entire $j\omega$ axis. Let $[Z_1]_e$ and $[Z_2]_e$ be expressed by Gauss factorization[9] as

$$[Z_1]_e = NL^{-1}L_*^{-1}N'$$
$$[Z_2]_e = MJ^{-1}J_*^{-1}M'$$

where N and M are n x n even polynomial matrices and L and J are n x n diagonal Hurwitz polynomial matrices.

Then Z_1 and Z_2 are compatible if and only if there exists an n x n para-unitary matrix $\mathcal{E}$ such that with $P = L\mathcal{E}J^{-1}$

1. $Z_{22} = MP'[(JJ)^{-1}(P_o)^{-1}]'M' - Z_2$ is a reactance matrix,
2. $[N(MP' - Z_2M'^{-1}JJ_*P_o')^{-1}]_\infty$ exists, and
3. $[\frac{1}{s}Z_1]_\infty - [\frac{1}{s}NP'^{-1}(PM'Z_2^{-1}M - P_oJJ_*)^{-1}N']_\infty$ is positive semi-definite if Z_{22} has a pole at infinity. Furthermore, the lossless coupling network is characterized by the following impedance matrix $\hat{Z}$:

$$\hat{Z} = \begin{bmatrix} NP'^{-1}(JJ_*)^{-1}(P_o)^{-1}N' + Z_1 & N[(JJ_*)^{-1}(P_o)^{-1}]'M' \\ M(JJ_*)^{-1}(P_o)^{-1}N' & MP'[(JJ_*)^{-1}(P_o)^{-1}]'M' - Z_2 \end{bmatrix}$$

(In the above expressions, the superscript prime and subscript asterisk stand for transposition and Hurwitzian conjugation, respectively. $[\ ]_\infty$ stands for the bracketed expression evaluated for $s \to \infty$ and subscripts e and o for even and odd parts.)

Since the proofs of these theorems are fairly long, they are not presented here; the reader should consult reference 10 for details. However, the outline of the proof for Theorem II is presented in the next section and illustrated with an example.

TRANSMISSION MATRIX APPROACH

Consider the situation depicted in Fig. 2, which illustrates the general case of compatibility. If A, B, C, and D are the generalized transmission parameters of $\hat{N}$, then the following relation exists between these parameters and the open circuit impedance parameters of $\hat{N}$:

$$\hat{Z} = \begin{array}{cc} \begin{array}{cc} \;\; n \;\; & \;\; n \;\; \end{array} & \\ \left[\begin{array}{c|c} Z_{11} & Z_{12} \\ \hline Z'_{12} & Z_{22} \end{array}\right] & \begin{array}{c} n \\ n \end{array} \end{array} = \left[\begin{array}{cc} AC^{-1} & C'^{-1} \\ C^{-1} & C^{-1}D \end{array}\right] \tag{1}$$

Since $\hat{Z}$ is a purely reactive network, A and D are even, and B and C are odd matrices of rational functions in s.

Writing Z_1 in terms of the transmission parameters and the termination matrix Z_n, one has

$$Z_1 = [AZ_n + B]\,[CZ_n + D]^{-1} \tag{2}$$

Let K be an n-order auxiliary rational matrix such that

$$CZ_n + D = K\ , \tag{3}$$

$$AZ_n + B = Z_1K\ . \tag{4}$$

Equating even and odd parts of the above expression and solving for the ABCD parameters, we get

$$A = [(Z_1)_e K_e + (Z_1)_o K_o]\,[(Z_n)_e]^{-1}\ , \tag{5}$$

$$B = [(Z_1)_e K_o + (Z_1)_o K_e] - [(Z_1)_e K_e + (Z_1)_o K_o]\,[(Z_n)_e]^{-1}[(Z_n)_o]\ , \tag{6}$$

$$C = K_o\,[(Z_n)_e]^{-1}\ , \tag{7}$$

$$D = K_e - K_o\,[(Z_n)_e]^{-1}\,[(Z_n)_o]\ . \tag{8}$$

Since the lossless network is assumed to be reciprocal, A, B, C, and D are n x n matrices satisfying two kinds of identities:

$$\left.\begin{array}{l} A'C - C'A \equiv 0_n \\ D'B - BD' \equiv 0_n \\ DC' - CD' \equiv 0_n \\ AB' - BA' \equiv 0_n \end{array}\right\} \qquad \left.\begin{array}{l} AD' - BC' \equiv 1_n \\ DA' - B'C \equiv 1_n \end{array}\right\} \tag{9, 10}$$

It can be shown that the above restrictions, arising from the generalization of quadripole formulae and first derived by Collet,[11] can be replaced by the following single equality:

$$K' [Z_1]_e K_* = [Z_n]_e , \tag{11}$$

or

$$[Z_1]_e K_* [(Z_n)_e]^{-1} K' = 1_n . \tag{12}$$

Now K can be determined from either of the two conditions:

$$[Z_1]_e K_* [Z_n]_e^{-1} K' = 1_n ,$$

or

$$K'[Z_1]_e K_* = [Z_n]_e .$$

Since both $[Z_1]_e$ and $[Z_n]_e$ are valid even parts of pr matrices, they can be factored by the Gauss factorization scheme. Thus, we obtain

$$[Z_1]_e = NL^{-1} L_*^{-1} N_*' , \tag{13}$$

$$[Z_n]_e = MJ^{-1} J_*^{-1} M_*' . \tag{14}$$

Since Z_1 and Z_n are symmetric, we have $N = N_*$ and $M = M_*$, i.e., N and M are n x n even polynomial matrices. L and J are n x n diagonal matrices of strictly Hurwitz polynomials. Substituting $[Z_1]_e$ and $[Z_n]_e$ from Eqs. (13) and (14) into (11) yields

$$K'NL^{-1}L_*^{-1}N'K_* = MJ^{-1}J_*^{-1} M' . \tag{15}$$

One solution therefore is

$$K' NL^{-1} = MJ^{-1}. \tag{16}$$

Because $[Z_1]_e$ is assumed to be non-singular, N is also non-singular, and one can write

$$K = N'^{-1} LJ^{-1} M' . \tag{17}$$

A general expression for K is obtained as follows: Let

$$[Z_1]_e = NL^{-1}L_*^{-1}N' = N\Lambda'^{-1} \Lambda_*^{-1} N' \tag{18}$$

$$[Z_n]_e = MJ^{-1}J_*^{-1}M' = M\Delta'^{-1} \Delta_*^{-1} M' , \tag{19}$$

where Λ and Δ are related to the diagonal matrices L and J by n x n regular para-unitary matrices Θ and Φ, i.e., $\Lambda = L\Theta_*'$, and $\Delta = J\Phi_*'$. Then K is given by

$$K = N'^{-1}\Lambda \Delta^{-1}M' = N'^{-1} L \mathcal{E} J^{-1}M' , \tag{20}$$

where the matrix $\mathcal{E} = [\Phi^{-1} \Theta]_*'$ is again seen to be para-unitary. Let

$$L\mathcal{E} J^{-1} = P = \Lambda \Delta^{-1} . \tag{21}$$

Then, we have

$$K = N'^{-1} PM' , \tag{22}$$

and, therefore, it can be written that

$$K_e = N'^{-1} P_e M' , \tag{23}$$

$$K_o = N'^{-1} P_o M' . \tag{24}$$

Substituting these expressions for K_e and K_o in Eqs. (5) to (8), one derives

$$A = [(Z_1)_e N'^{-1} P_e M' + (Z_1)_o N'^{-1} P_o M'] \, [(Z_n)_e]^{-1}, \tag{25}$$

$$B = [(Z_1)_e N'^{-1} P_o M' + (Z_1)_o N'^{-1} P_e M'] - [(Z_1)_e N'^{-1} P_e M' + (Z_1)_o N'^{-1} P_o M'] \, [(Z_n)_e]^{-1} \, [Z_n]_o, \tag{26}$$

$$C = N'^{-1} P_o M' \, [(Z_n)_e]^{-1}, \tag{27}$$

$$D = N'^{-1} P_e M' - N'^{-1} P_o M' \, [(Z_n)_e]^{-1} \, [Z_n]_o . \tag{28}$$

Hence, the parameters of the Z matrix are given by

$$Z_{11} = AC^{-1} = NP'^{-1} \, [JJ_*]^{-1} \, [P_o]^{-1} N' + Z_1 , \tag{29}$$

$$Z_{12} = C'^{-1} = N \, [(JJ_*)^{-1} \, (P_o)^{-1}]' M' , \tag{30}$$

$$Z_{22} = C^{-1} D = MP' \, [(JJ_*)^{-1} \, (P_o)^{-1}]' M' - Z_n . \tag{31}$$

Next, the realizability of $\hat{Z}$ is considered. Clearly $\hat{Z}$ is lossless and symmetric by construction, a fact which is expressed by the relation $\hat{Z} = -\hat{Z}'_*$. To prove that $\hat{Z}$ as constructed above is a pr matrix under certain conditions, one applies the well-known test for its positive real nature. It can be shown that the conditions given in the theorem are indeed the ones that make $\hat{Z}$ a pr matrix.

EXAMPLE

To illustrate the above theorem, let us determine if the following two impedance matrices are compatible:

$$Z_1 = \frac{s}{(2s+1)(3s+1)} \begin{bmatrix} 4s^2+10s+3 & 2s^2+6s+2 \\ 2s^2+6s+2 & 7s^2+10s+3 \end{bmatrix},$$

$$Z_n = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} .$$

Computing even parts yields

$$[Z_1]_e = \frac{1}{dd_*} \begin{bmatrix} 5s^2(8s^2-1) & 2s^2(13s^2-2) \\ 2s^2(13s^2-2) & 5s^2(5s^2-1) \end{bmatrix},$$

where $d = 6s^2 + 5s+1 = (2s+1)(3s+1)$. Also, we have $[Z_n]_e = [Z_n]$. With $[Z_1]_e$ and $[Z_n]_e$ expressed in the standard form, one obtains

$$[Z_1]_e = NL^{-1} L_*^{-1} N' ,$$

$$[Z_n]_e = MJ^{-1} J_*^{-1} M' ,$$

where, for this example,

$$N = \begin{bmatrix} 5s^2(8s^2-1) & 0 \\ 2s^2(13s^2-2) & 9s^2 dd_* \end{bmatrix}$$

$$L = \begin{bmatrix} \sqrt{5}\, s(\sqrt{8}\, s+1)d & 0 \\ 0 & \sqrt{5}\, s(\sqrt{8}\, s+1)d^2 \end{bmatrix},$$

$$M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

$$J = 1_2 .$$

It can now be shown that the following regular para-unitary matrix $\mathcal{E}$ is such that Z_{22} is indeed a reactance matrix:

$$\mathcal{E} = \begin{bmatrix} \dfrac{(2s-1)}{\sqrt{5}(\sqrt{8}s+1)} & \dfrac{2(3s-1)}{\sqrt{5}(\sqrt{8}s+1)} \\ \dfrac{2(2s-1)(3s-1)}{\sqrt{5}(\sqrt{8}\, s+1)(2s+1)} & -\dfrac{(2s-1)(3s-1)}{\sqrt{5}(\sqrt{8}\, s+1)(3s+1)} \end{bmatrix}.$$

Therefore, we have

$$P = L\mathcal{E}J^{-1} = \begin{bmatrix} s(4s^2-1)(3s+1) & 2s(9s^2-1)(2s+1) \\ 6s(4s^2-1)(9s^2-1)(3s+1) & -3s(9s^2-1)(4s^2-1)(2s+1) \end{bmatrix}.$$

Taking the odd part of P, one finds that

$$P_o = \begin{bmatrix} 4(4s^2-1) & 2s(9s^2-1) \\ 6s(4s^2-1)(9s^2-1) & -3s(9s^2-1)(4s^2-1) \end{bmatrix}.$$

Substituting the respective matrices for M, J, P, P_o, and Z_n in Eq. (31) and simplifying, we obtain,

$$Z_{22} = \begin{bmatrix} 3s & 3s \\ 3s & 5s \end{bmatrix}.$$

Since Z_{22} has a pole at infinity, it is also necessary to test for conditions 2 and 3 of Theorem 4. Routine calculation shows that these conditions are indeed satisfied. Hence, Z_1 is compatible to Z_n, and the lossless coupling structure is characterized by

$$\hat{Z} = s \begin{bmatrix} 3 & 2 & 1 & 3 \\ 2 & 3 & 2 & 3 \\ 1 & 2 & 3 & 3 \\ 3 & 3 & 3 & 5 \end{bmatrix}.$$

DISCUSSION OF THE RESULTS

The present investigation has extended the notion of compatibility of driving-point immittances to immittance matrices. Two quite general theorems, giving the necessary and sufficient conditions for the compatibility of immittance matrices, were presented. These theorems, however, are not explicit, and a serious drawback in their application stems from the necessity of establishing the existence of an unknown para-unitary matrix. This matrix, of course, is not arbitrary and must satisfy certain restrictions arising from the physical realizability and losslessness of the coupling structure and from given immittance matrices whose compatibility is under investigation. This situation arises because of the non-uniqueness involved in the factorization of para-Hermitian positive semi-definite matrices, a step which is crucial in the synthesis of the coupling structure. This non-uniqueness is simply a manifestation of the fact that many equivalent coupling structures are capable of making a given pair of immittance matrices compatible. By considering a specific class of terminations and by imposing topological constraints on the coupling structure, this non-uniqueness can, in some cases, be avoided.[12]

APPLICATIONS

Two useful applications of these results deserve mention. The first important application arises in the problem of optimum matching of a source with arbitrary internal impedance $z_2(s)$ to an arbitrary load $z_1(s)$. If z_1 and z_2 are compatible, then the scheme suggested by Fig. 4 achieves an optimum power match between the source and the load. This fact is obvious since the coupling network $\hat{N}$ shown in the dotted enclosure is lossless. Rohrer[13] has shown that if z_1 is minimum reactive, it is always possible to find an impedance $z'(s)$ such that $-2(z_1)_o = (z')_o$.

The synthesis of the odd part of a pr function is always possible with the help of a lossless passive network and a negative impedance convertor, as shown by Mitra.[14] By utilizing the results of the present investigation, this technique can easily be extended to handle the problem of multi-port matching.

Secondly, the notion of compatibility is intimately related to that of lossless reciprocal imbedding. Referring to Fig. 5, one sees that Z_1 is obtained by lossless reciprocal impedding of Z_2. Mason[15, 16] points out that the set of all lossless reciprocal transformations forms a group and that any such imbedding is a result of three simple lossless transformations; namely, (1) addition of a reactance, (2) a congruence transformation, and (3) inversion. If the operation is confined to a single real frequency, it can easily be shown that (1) Z is always compatible to Z^{-1}, and vice versa, (2) the compatibility is mutual, i.e., if Z_1 is compatible to Z_2, then Z_2 is also compatible to Z_1, and vice versa.

The present study also leads to the following generalization of the compatibility concept as suggested by DeClaris. Suppose, for example, that the two pr impedances z_1 and z_2 are not compatible in the sense that compatibility is defined here. Now it is asked, can z_1 be realized as a passive two-port terminated by z_2? Thus the coupling network is freed from the imposition of being lossless.

Since a general passive two-port can always be realized as a lossless four-port terminated in two unit resistances at most (Fig. 5), the problem essentially reduces itself to the study of the compatibility (in the usual sense) of two immittance matrices

Z_1 and Z_2, with Z_1 being of smaller order than Z_2. This situation has an immediate relevance to the problem of establishing an optimal power match between number of loads and a common source feeding them.

Thus, in conclusion, the most useful extension of the present work would be in making the compatibility conditions explicit and in deriving results applicable to the situation where the terminating ports are more numerous than the input ports.

ACKNOWLEDGMENTS

The author gratefully acknowledges the help and guidance given him by Dr. S.K. Mitra and Prof. N. DeClaris of the School of Electrical Engineering, Cornell University.

REFERENCES

1. J.D. Schoeffler, "Impedance Transformation Using Lossless Networks," IEEE Trans. on Circuit Theory, Vol. CT-8, pp. 131-137, June, 1961.
2. D.C. Youla, "An Extension of the Concept of Scattering Matrix," IEEE Trans. on Circuit Theory, Vol. CT-11, pp. 310-312, June, 1964.
3. R.A. Rohrer, "The Scattering Matrix Normalized to Complex n-Port Load Networks," Report R-206, Coordinated Science Lab., Univ. of Illinois, May, 1964.
4. M.R. Wohlers, "Complex Normalization of Scattering Matrices and the Problem of Compatible Impedances," IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 528-536, December, 1965.
5. M. Bayard, "Resolution du Probleme de la Synthese des Reseaux de Kirchoff par la Determination de Reseaux Puremen Reactifs," Cables et Transmission, pp. 281-296, October, 1950.
6. D.C. Youla, "On the Factorization of Rational Matrices," IEEE Trans. on Information Theory, IT-7, pp. 172-189, July, 1961.
7. R.W. Brockett, "Spectral Factorization of Rational Matrices," to be published. For summary, see Int. Conf. on Microwaves, Circuit Theory and Information Theory, Summaries of Papers, Part 3, Tokyo: Inst. of Electrical Communication Engineers of Japan, (1964).
8. V. Belevitch, "Factorization of Scattering Matrices with Applications to Passive Network Synthesis," Phillips Res. Rep., Vol. 18, pp. 275-317, August, 1963.
9. R.W. Newcomb, "A Bayard Type Nonreciprocal n-Port Synthesis," IEEE Trans. on Circuit Theory, CT-10, pp. 85-90, March, 1963.
10. S.K. Mullick, "Compatibility of Immittance Matrices," Research Report EERL 37, Cornell University, Ithaca, June, 1965.
11. L. Collet, "Definition et Carateristiques des Circuits Superposes," Ann. telecomm., Vol. 4, pp. 42-56, February, 1949.
12. I.W. Sandberg, "Realizability of Multi-port Structures Obtained by Imbedding a Tunnel Diode in a Lossless Reciprocal Network," Bell System Tech. J., Vol. 41, pp. 857-876, May, 1962.

13. R.A. Rohrer, "Active Network Matching of Arbitrary Loads," Series No. 60, Issue No. 367, Electronics Res. Lab., Univ. of Calif., Berkeley, June, 1961.

14. S.K. Mitra, "A New Approach to Active RC Network Synthesis," J. Franklin Inst., Vol. 274, pp. 185-197, September, 1962.

15. S. Mason, "Power Gain in Feedback Amplifiers," IEEE Trans. on Circuit Theory, CT-1, pp. 20-25, June, 1954.

16. H.J. Carlin and A.B. Giordano, "Network Theory," Prentice-Hall, Englewood Cliffs, N.J., 1964.

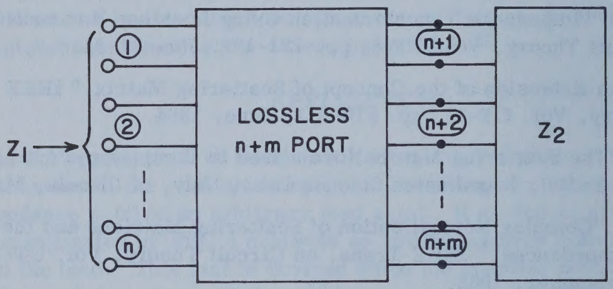

Z₁ →
①
②
ⓝ
LOSSLESS
n+m PORT
(n+1)
(n+2)
(n+m)
Z₂

Fig. 1. Compatibility of Two Impedance Matrices.

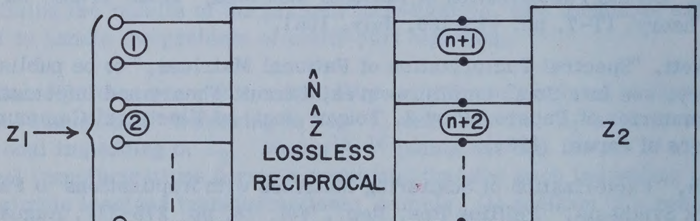

Z₁ →
①
②
ⓝ
N̂
Ẑ
LOSSLESS
RECIPROCAL
(n+1)
(n+2)
(2n)
Z₂

Fig. 2. Compatibility of Z_1 and Z_2, both Being the Same Order.

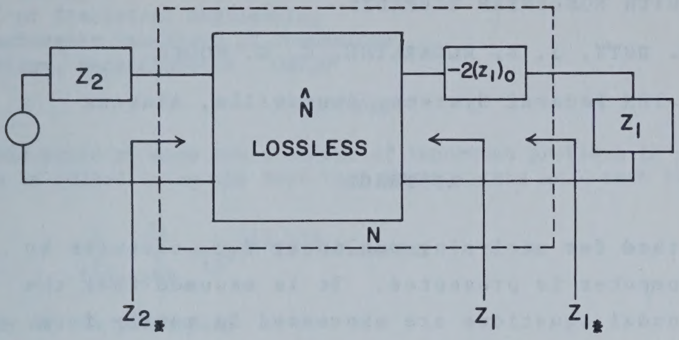

Z_2
$\hat{N}$
LOSSLESS
$-2(z_1)_0$
Z_1
N
Z_{2*}
Z_1
Z_{1*}

Fig. 3. Optimum Power Match Between Source and Load.

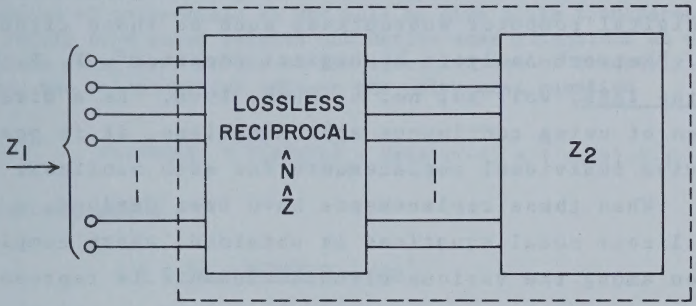

Z_1
LOSSLESS
RECIPROCAL
$\hat{N}$
$\hat{Z}$
Z_2

Fig. 4. Lossless Reciprocal Imbedding of Z_2.

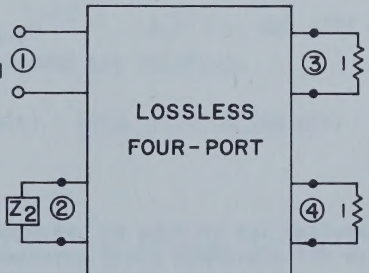

Z_1 ①
③ 1
LOSSLESS
FOUR-PORT
Z_2 ②
④ 1

Fig. 5. Compatibility with a Passive, Not Necessarily Lossless, Network.

A NUMERICAL METHOD FOR DETERMINING NODAL VOLTAGES IN D.C. CIRCUITS WITH NONLINEAR ELEMENTS

R. L. DUTY, J. B. EGGERLING, C. G. HOOKS

IBM Federal Systems, Huntsville, Alabama

ABSTRACT

A method for analyzing nonlinear d.c. circuits by digital computer is presented. It is assumed that the original nodal equations are expressed in matrix form. The method uses a sequence of continuous approximations to the characteristic curve for each nonlinear element, and thus certain convergence problems introduced by piecewise--linear approximations are avoided. The method is compatible with existing digital computer subroutines such as those cited in the paper, "Network Analysis by Digital Computer", F. F. Kuo, Proc. of the IEEE, vol. 54, no. 6, June, 1966. As a direct consequence of using continuous approximations, it is possible to derive individual replacements for each nonlinear element. When these replacements have been derived, a new system of linear nodal equations is obtained, where complete interaction among the various circuit elements is represented. This system of equations is iteratively solved to obtain the final nodal voltages. The general procedure is illustrated by application to several nonlinear circuits. Numerical results obtained on an IBM 360/50 computer are presented to show the rates of convergence.

PATH INTEGRALS, LIAPUNOV FUNCTIONS, AND QUADRATIC MINIMIZATION*

R. W. BROCKETT
Dept. of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

In this paper we show how a number of important problems in system theory can be solved using the fact that there exists an η such that

$$\sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}\phi^{(i)}\phi^{(j)} = d\eta/dt$$

if and only if the polynomial

$$h(s) = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}[s^{(i)}(-s)^{j} + s^{(j)}(-s)^{(i)}]$$

vanishes identically. Using this result and elementary facts about the factorization of polynomials we are able to give quite transparent proofs of some fairly hard known results and derive some extensions as well. The applications considered relate to the effective use of frequency domain ideas and, the relationships between the polynomial equation

$$r(s)p(-s) + r(-s)p(s) + \beta r(s)r(-s) = + \alpha q(s)q(-s)$$

and its matrix analog

$$\underline{A}'\underline{K} + \underline{KA} - \beta\underline{Kbb'K} = - \alpha\underline{cc}'$$

NOTATION

Lower case Latin and Greek letters with underlines are column vectors, underlined capitals denote square matrices. A prime will denote transpose. We let C^n denote the class of all n times differentiable functions of time. If x belongs to C^n then $x^{(i)}$ denotes its ith derivative with respect to time. We will also use an operator notation $D^i x = x^{(i)}$ and even $\dot{x} = x^{(1)}$

If $p(D) = D^n + p_{n-1}D^{n-1} + \ldots p_1 D + p_0$ and $q(D) = q_{n-1}D^{n-1} + q_{n-2}D^{n-2} + \ldots q_1 D + q_0$, then clearly the equations

$$p(D)x(t) = u(t); \quad y(t) = q(D)x(t) \tag{1}$$

---

*This work was supported in part by the National Aeronautics and Space Agency through NASA Research Grant NGR22-009-124 with the M.I.T. Electronics Systems Laboratory.

and

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}(t) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & 1 \\ -p_0 & -p_1 & -p_2 & \dots & -p_{n-1} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}(t) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}u(t) \tag{2}$$

$$y(t) = (q_0, q_1, \dots q_{n-2}, q_{n-1})\,(x_1, x_2 \dots x_{n-1}, x_n)'(t) \tag{3}$$

are equivalent and $x_i = x^{(i-1)}$. The transfer function in both cases is

$$g(s) = q(s)/p(s) \tag{4}$$

PATH INTEGRALS

Suppose that ϕ belongs to C^n and that α_{ij} are constants. The integral

$$\eta = \int_a^b \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij}\phi^{(i)}(t)\phi^{(j)}(t)dt \tag{5}$$

is then unambiguously defined. In general its value will depend not only on the value taken by ϕ and its derivatives at a and b but also on their values between a and b. However if the integrand in equation (5) is for example $\dot{\phi}(t)\phi(t)$ then clearly $\eta = \frac{1}{2}[\phi^2(b) - \phi^2(a)]$ regardless of the form of ϕ. We propose the following definition.

Definition: The integral η defined by equation (5) will be said to be independent of path if it can be evaluated in terms of ϕ and its first n-1 derivatives at a and b.

Integrals which are independent of path play an important role here. The standard results on path independence apply to the situation

$$\eta_1 = \int_{\underline{a}}^{\underline{b}} \sum_{i=1}^{n} f_i(\underline{x})dx_i \tag{6}$$

and require (in addition to smoothness conditions) that

$$\partial f_i/\partial x_j = \partial f_j/\partial x_i \tag{7}$$

for path independence. (See reference 1). It is true that one can suppress the time dependence in equation (5) to get an integral of the form ($\underline{a}$ and $\underline{b}$ not related to those above)

$$\eta = \int_{\underline{a}}^{\underline{b}} \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}x_i dx_j; \quad x_i = \phi^{(i-1)} \tag{8}$$

But, in this form the symmetry condition (7) only gives an unnecessarily restrictive sufficient condition because of the nonintegrable constraint equations

$$\frac{dx_1}{x_2} = \frac{dx_2}{x_3} = \frac{dx_{n-1}}{x_n} = dt \tag{9}$$

which constrain the paths of integration.

TWO LEMMAS ON PATH INTEGRALS

In this section we state and prove our main results. Lemma 1 is a generalization of Lemma 1 in reference [2]. The method of proof is identical. Wonham and Johnson [3] have also used (implicitly) a special case.

Lemma 1: Assume that ϕ belongs to C^n and that α_{ij} are constants. The integral (5) is independent of path if and only if the polynomial

$$h(s) = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}[s^i(-s)^j + (-s)^i s^j] \tag{10}$$

vanishes identically.

Proof: Notice that if i + j is odd then by using the integration-by-parts formula

$$\int_a^b \phi^{(\nu)}(t)\phi^{(\mu)}(t)dt = \phi^{(\nu-1)}(t)\phi^{(\mu)}(t)\Big|_a^b - \int_a^b \phi^{(\nu-1)}(t)\phi^{(\mu+1)}(t)dt \tag{11}$$

a total of $|i - j| -1$ times the integral is reduced to one of the form $x^{(p)}(t)x^{(p-1)}(t)$ which can be integrated. If i + j is even the problem is more complicated. Using equation (11) a total of $|i - j|$ times leaves one with $|i - j|$ terms expressed in terms of the end points plus a term $(-1)^{(i-j)}\{x^{([i + j]/2)}\}^2$ inside the integral. Adding all these up gives

$$\eta = \eta_0 + \int_a^b \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}[(-1)^i + (-1)^j]\{\phi^{([i+j]/2)}\}^2 dt \tag{12}$$

where η_0 is a function of the end points only and the sum is over i + j even. But since

$$\sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}[s^i(-s)^j + (-s)^i s^j] = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}[(-1)^j + (-1)^i]s^{i+j} = 0 \tag{13}$$

We see that the integral term in equation (12) vanishes if and only if equation (10) holds. This completes the proof.

Now it is clear that if the integrability condition (10) holds then η can be expressed as

$$\eta = \int_a^b \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}\phi^{(i)}(t)\phi^{(j)}(t)dt =$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij}\phi^{(i-1)}(t)\phi^{(j-1)}(t)\Big|_a^b \tag{14}$$

with $k_{ij} = k_{ji}$. That is to say, if η is expressible in terms of the end points it is a quadratic form in ϕ and its first n-1 derivatives. Since the times a and b play no role in our applications we introduce the notation

$$\eta = \int_{t(\underline{0})}^{t(\underline{x})} \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij}\phi^{(i)}\phi^{(j)}dt = \underline{x}'\underline{K}\underline{x} \tag{15}$$

to describe a path integral which starts at $\underline{0}$ (i.e. $\phi[t(\underline{0})] = \phi^{(1)}[t(\underline{0})] = \ldots \phi^{(n-1)}[t(\underline{0})] = 0$) and ends at a point $\underline{x}$ (i.e. $\phi[t(\underline{x})] = x_1$, $\phi^{(1)}[t(\underline{x})] = x_2 \ldots \phi^{(n-1)}[t(\underline{x})] = x_n$). Thus this integral is to be interpreted as a line integral in the state space (see Figure 1).

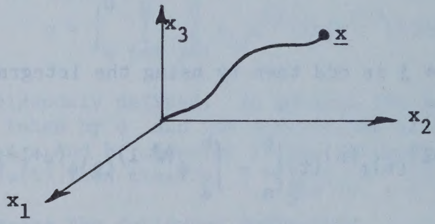


Figure 1. Illustrating the meaning of $\eta(\underline{x})$.

Equation (15) is nothing more than a integral representation for a Quadratic form just as is

$$\eta(\underline{x}) = \int_0^\infty \underline{x}'e^{\underline{A}'t}\underline{C}e^{\underline{A}t}\underline{x}dt \tag{16}$$

Its importance stems from the fact that $\underline{x}'\underline{K}\dot{\underline{x}}$ for such systems can be calculated simply by removing the integral sign. This is a consequence of the following lemma.

Lemma 2: Consider $\eta(\underline{x})$ as defined by the path independent integral (15). If $\alpha_{ij} = \alpha_{ji}$, as we can assume without loss of generality, then $\underline{K} = \underline{K}'$ is zero if and only if all the α_{ij} are zero.

Proof: If all the α_{ij} are zero then clearly $\underline{K} = \underline{0}$. Find the terms for which i + j is a maximum and from these select the term for which i is a maximum. If this term is $\alpha_{\nu\mu}$ then from our evaluation of $\underline{K}$ it follows that $k_{\nu(\mu+1)} = \alpha_{\nu\mu}$ and hence $\underline{K}$ cannot be zero if any of the α_{ij} are nonzero.

In preparation of latter applications we observe that with our previous definitions of p(D) and q(D), the integral

$$\eta = \int_a^b p(D)\phi(t)q(D)\phi(t) - [r(D)\phi(t)]^2 dt \tag{17}$$

is independent of path if

$$\tfrac{1}{2}[p(D)q(-D) + p(D)q(-D)] = r(D)r(-D) \tag{18}$$

If this condition holds then we can write

$$\underline{x}'\underline{Kx} = \int_{t(\underline{0})}^{t(\underline{x})} p(D)\phi(t)q(D)\phi(t) - [r(D)\phi(t)]^2 dt \tag{19}$$

and

$$2\underline{x}'\underline{K}\dot{\underline{x}} = p(D)x_1 q(D)x_1 - [r(D)x_1]^2 \tag{20}$$

Along solutions of the differential equations

$$p(D)x(t) = 0 \tag{21}$$

we have,

$$\frac{d}{dt}\,\underline{x}'(t)\underline{Kx}(t) = -\,[r(D)x(t)]^2 \tag{22}$$

APPLICATION TO LEAST SQUARES PROBLEMS

Consider equation (2) together with the problem of minimizing the performance measure ($\underline{Q}$ constant and positive semidefinite)

$$\eta_0 = \int_a^b \underline{x}'(t)\underline{Qx}(t) + \alpha u^2(t)dt; \quad b > a \tag{23}$$

Consider the following.

Problem 1: Given equation (2) and given fixed end points $\underline{x}(a)$ and $\underline{x}(b)$, or the performance measure η_0, find a constant vector $\underline{c}$ such that the performance measure

$$\eta_1 = \int_a^b [\underline{c}'\underline{x}(t)]^2 + \alpha u^2 dt \tag{24}$$

gives rise to the same minimizing control.

Solution: Let $j = \sqrt{-1}$ and let $\underline{\omega} = \text{col}(1, j\omega, \ldots (j\omega)^{n-1})$ Then if $\underline{\omega}^*$ denotes conjugate transpose, $\underline{\omega}^*\underline{Q\omega}$ is real and nonnegative for all real ω. Hence there exists real constants $c_1, c_2, \ldots c_n$ (not unique) such that

$$\underline{\omega}^*\underline{Q\omega} = |c_1 + c_2 j\omega + \ldots c_n (j\omega)^{n-1}|^2 \tag{25}$$

That is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}D^{i-1}(-D)^{j-1} = c(D)c(-D) \tag{26}$$

with

$$c(D) = c_1 + c_2D + \ldots c_nD^{n-1} \tag{27}$$

From Lemma 1 the integral

$$\eta_2 = \int_a^b - \underline{x}'(t)\underline{Q}\underline{x}(t) + [\underline{c}'\underline{x}(t)]^2dt \tag{28}$$

is independent of path. Since the end points $\underline{x}(a)$ and $\underline{x}(b)$ in our problem are fixed we may add η_2 to η_0 without affecting the variational problem. But since

$$\eta_0 + \eta_2 = \int_a^b [\underline{c}'\underline{x}(t)]^2 + \alpha u^2(t)dt \tag{29}$$

we see that this $\underline{c}$ is of the form asked for in the problem statement.

Notice that this approach will work for time varying systems (but not time varying $\underline{Q}$'s). This is the type of application Wonham and Johnson[3] describe. The special case $a = 0$, $b = \infty$ plays an important role in Kalman's paper [4] but the method of proof used there is much less direct.

As a more important application we consider the following

Problem 2: Consider equations (2) and (3). Given $\underline{x}(0)$ find $u(t)$ such that $\underline{x}(t)$ approaches zero as t approaches infinity and

$$\eta_3 = \int_0^\infty \alpha y^2(t) + u^2(t)dt \tag{30}$$

is a minimum. Assume the system is observable.

Solution: We will solve this problem directly without appealing in any way to the calculus of variations. Use the definition of equation (1) to write η_3 as

$$\eta_3 = \int_0^\infty \alpha[q(D)x(t)]^2 + (p(D)x(t)^2dt \tag{31}$$

If $r(D)$ is any solution of

$$\alpha q(D)q(-D) + p(D)p(-D) = r(D)r(-D) \tag{32}$$

then by lemma 1

$$\eta_4 = \int_0^\infty - [p(D)x(t)]^2 - \alpha[q(D)x(t)]^2 + [r(D)x(t)]^2dt \tag{33}$$

depends only on x and its first $n-1$ derivatives at zero and infinity.

Since these are both fixed by the problem statement we may add η_4 to η_3 without affecting the form of the minimizing u(t). But

$$\eta_3 + \eta_4 = \int_0^\infty [r(D)x(t)]^2 dt \tag{34}$$

Clearly we want to pick x so that r(D)x(t) = 0 provided we can do so and still meet our boundary condition at infinity. But as long as α is positive equation (32) has a solution for r(D) which has its zeros in Re $s < 0$ (absence of common factors of q(D) and p(D) used here). The usual notation for this (spectral) factor is

$$r(D) = [p(D)p(-D) + \alpha q(D)q(-D)]^+ \tag{35}$$

Clearly u(t) is given in terms of x(t) as

$$u(t) = p(D)x(t) - r(D)x(t) \text{ so } h(D)x(t) = 0 \tag{36}$$

Notice that the minimizing value of u(t) gives

$$\eta_3 = \int_{t(\underline{0})}^{t(\underline{x})} \{[p(D)p(-D) + \alpha q(D)q(-D)]^+ \phi(t)\}^2 - [p(D)\phi(t)]^2 - \alpha[q(D)\phi(t)]^2 dt \tag{37}$$

Of course both the spectral factorization solution and explicit expressions for η_3 have been known for some time. The point is that this derivation requires no use of either transform methods or calculus of variations.

Finally, Kalman (e.g. reference [4]) has shown that the value of η_3 is $\underline{x}'\underline{Kx}$ where $\underline{K}$ satisfies the quadratic matrix equation

$$\underline{A}'\underline{K} + \underline{KA} - \underline{KbbK} = -\alpha\underline{cc}' \tag{38}$$

and is positive definite. Thus we see that if $\underline{A}$ and $\underline{b}$ have the form given by equation (2) and if $\underline{K}$ is the positive definite solution of equation (38) then

$$\underline{x}'\underline{Kx} = \int_{t(\underline{0})}^{t(\underline{x})} \{[p(D)p(-D) + \alpha q(D)q(-D)]^+ \phi(t)\}^2 - [p(D)\phi(t)]^2 - \alpha[q(D)\phi(t)]^2 dt \tag{39}$$

By a change of notation whereby $\underline{A}$ is replaced by $\underline{A}$-$\underline{bc}'$ we can get the solution to a more general form of equation (38). In fact, changing $\underline{A}$ to $\underline{A}$-$\underline{bc}'$ merely replaces p(D) by p(D) + q(D) in equation (39). Notice that in this case the choice $\alpha = -1$ gives the integral which was used in reference [2] to derive the Popov criterion.

APPLICATION TO STABILITY PROBLEMS

Lemma 1 makes it possible to derive in a simple way several different forms of Hermite-type inequalities which give necessary and sufficient conditions for all the zeros of a polynomial with real coefficients to have its zeros in the half-plane Re $s < 0$.

Problem 3 : Derive necessary and sufficient conditions for a polynomial with real coefficients

$$p(s) = s^n + p_{n-1}s^{n-1} + \cdot \cdot \cdot p_1 s + p_0 \tag{40}$$

to have all its zeros in the half-plane Re $s < 0$.

Solution : Since $p(s)p(-s) - p(-s)p(s)$ vanishes identically we can write without ambiguity

$$\underline{x}'\underline{Kx} = \int_{t(\underline{0})}^{t(\underline{x})} [p(D)\phi(t)]^2 - [p(-D)\phi(t)]^2 dt \tag{41}$$

The derivative of $\underline{x}'\underline{Kx}$ along solutions of $p(D)x(t) = 0$ is,

$$\frac{d}{dt}\,\underline{x}'\underline{Kx} = -[p\,(-D)x(t)]^2 \tag{42}$$

Clearly $p(-D)$ and $p(D)$ do not have common factors if the zeros of $p(s)$ lie in the half-plane Re $s < 0$. Thus $p(-D)x(t)$ will not vanish identically along any nontrivial solution of $p(D)x(t) = 0$. Hence positive definiteness of $\underline{K}$ is sufficient to insure that all the zeros of $p(s)$ lie in Re $s < 0$. To show that it is necessary, assume that it is not positive definite but that the zeros of $p(s)$ do lie in Re $s < 0$. Then for some $\underline{x}$, $\underline{x}'\underline{Kx} = 0$. From equation (42) we see that $\underline{x}'\underline{Kx}$ is nonincreasing and since it must approach zero we see it is a constant. Thus we have a contradiction because $p(D)$ and $p(-D)$ cannot have common factors and hence $p(D)x(t) = 0$ and $p(-D)x(t) = 0$ have no nontrivial simultaneous solutions.

For the third degree polynomial $s^3 + as^2 + bs + c$ the matrix $\underline{K}$ takes the form

$$\underline{K}_3 = 2\begin{bmatrix} a & 0 & c \\ 0 & ab-c & 0 \\ c & 0 & bc \end{bmatrix} \tag{43}$$

As a second type of application consider proving stability for

$$\dot{\underline{x}} = \underline{Ax} - f(t)\underline{bc}'\underline{x} \tag{44}$$

with $\underline{A}$, $\underline{b}$ and $\underline{c}$ having the form given by equations (2) and (3) and $g(s)$ being positive real. It is known that there exists a time invariant quadratic Liapunov function whose derivative is negative semidefinite for all time, if $f(t) \geqslant 0$. However, this assumtion on $f(t)$ is more restrictive than necessary in many cases and one would like to replace the condition on $f(t)$ with one involving only "average values"; i.e. some integral condition on $f(t)$. Wazewski's theorem [5] allows one to do this but in order for this approach to give any improvement over Liapunov methods it is necessary that the Liapunov function be negative definite for some $f(t)$. Thus we are lead to consider the following

Problem 4 : Assume that the state vector in equations (2) and (3) is controllable and observable and that g(s) is a positive real function. Letting $u(t) = -f(t)y(t)$ in these equations gives equation (44). Find a constant, positive definite matrix $\underline{K}$ such that

$$\frac{d}{dt}\underline{x}'\underline{K}\underline{x} = \underline{x}'(\underline{K}\underline{A} + \underline{A}'\underline{K})\underline{x} - f(t)(\underline{c}'\underline{x})^2 \qquad (45)$$

and $\underline{K}\underline{A} + \underline{A}'\underline{K}$ is both negative semidefinite and of maximal rank.

Solution : Consider as a Liapunov function

$$\underline{x}'\underline{K}\underline{x} = \int_{t(\underline{0})}^{t(\underline{x})} p(D)\phi(t)q(D)\phi(t) - \sum_{i=1}^{v} [r_i(D)\phi(t)]^2 dt \qquad (46)$$

This definition makes sense if the path independence condition

$$\frac{1}{2}[p(D)q(-D) + q(D)p(-D)] = \sum_{i=1}^{v} r_i(D)r_i(-D) \qquad (47)$$

is satisfied. Moreover, since equation (44) is clearly equivalent to

$$p(D)x(t) + f(t)q(D)x(t) = 0 \qquad (48)$$

it follows that along solutions of equations (44)

$$\frac{d}{dt}\underline{x}'\underline{K}\underline{x} = -f(t)[q(D)x_1]^2 - \sum_{i=1}^{v} [r_i(D)x_1]^2 \qquad (49)$$

Thus the rank of $\underline{K}\underline{A} + \underline{A}'\underline{K}$ is just the number v. Thus the solution to our problem depends only on expressing the even polynomial

$$h(s) = p(s)q(-s) + q(s)p(-s) \qquad (50)$$

as a sum of a maximum number of squares. This number will be equal to one plus one half the degree of h(s) minus one half the number of zeros of h(s) on the imaginary axis. This argument really only gives a lower bound on the maximal rank of $\underline{K}\underline{A} + \underline{A}'\underline{K}$ but it can be shown that it cannot be improved upon.

APPLICATIONS IN NETWORK THEORY

Let u and y be scalars related by

$$\dot{\underline{x}}(t) = \underline{f}(\underline{x}(t), u(t)) \ ; \ y(t) = g(\underline{x}(t), u(t)) \qquad (51)$$

An intuitively pleasing definition for passivity for such systems is the following

Definition. The system (51) is said to be passive if there exists a positive definite function $v(\underline{x})$ along solutions of equation (51) is less than u y. That is,

$$\dot{v}(\underline{x}) = \sum_{i=0}^{n} (\partial v/\partial x_i)f_i(\underline{x},u) \leq u\ y \qquad (52)$$

We have in mind that $v(\underline{x})$ is the total energy of the system and $u\,y$ is the power delivered to the system. To support our definition we need to resolve the following

Problem 5. Given a linear constant system described by equations (2) and (3) show that if the state is both controllable and observable and if $g(s)/p(s)$ is positive real, then there exists a positive definite function $\underline{x}'\underline{K}\underline{x}$ such that

$$\underline{x}'(\underline{A}'\underline{K} + \underline{K}\underline{A})x + \underline{x}'\underline{K}\underline{b}u + \underline{b}'\underline{K}\underline{x}u \leq u\,y \tag{53}$$

Solution : For our function $\underline{x}'\underline{K}\underline{x}$ we pick the path independent integral.

$$\underline{x}'\underline{K}\underline{x} = \int_{t(\underline{0})}^{t(\underline{x})} p(D)\phi(t)q(D)\phi(t) - \{[\mathrm{Ev}\; p(D)q(-D)]^{+}\phi(t)\}^{2}dt \tag{54}$$

Using the equivalence between equation (1) and equation (2) and (3) we see that

$$\frac{d}{dt}\,\underline{x}'\underline{K}\underline{x} = u\,y - \{[\mathrm{Ev}\; p(D)q(-D)]^{+}x_{1}\}^{2} \tag{55}$$

which is certainly less than or equal to $u\,y$.

It remains only to show that $\underline{x}'\underline{K}\underline{x}$ is positive definite. To see this we consider evaluating $\underline{x}'\underline{K}\underline{x}$ along a particular set of paths, namely those for which

$$p(D)\phi(t) + q(D)\phi(t) = 0 \tag{56}$$

It is well known that $p(D) + q(D)$ has its zeros in the half-plane $\mathrm{Re}\, s < 0$ if $g(s)/p(s)$ is positive real. If we start at the state $\underline{x}$ then it will approach zero as $t \to \infty$. Thus

$$\underline{x}'\underline{K}\underline{x} = \int_{0}^{\infty} [q(D)\phi(t)]^{2} + \{[\mathrm{Ev}\; p(D)q(-D)]^{+}\phi(t)\}^{2}dt \tag{57}$$

but this is clearly positive definite because $q(D)x$ cannot vanish identically along any nonzero solution of equation (56) because $q(s)$ and $p(s)$ do not have common factors. Thus $\underline{x}'\underline{K}\underline{x}$ is positive definite.

As a second, and closely related application we can solve the following problem which is motivated by Darlington Synthesis.

Problem 6 : Show that if $g(s)$ for equation (2) and (3) is positive real then a suitable change of coördinates puts the equations in the form

$$\dot{\underline{z}} = \underline{S}\underline{z} - \underline{p}\underline{p}'\underline{z} + \underline{\gamma}u \;; y = \underline{\gamma}'\underline{z} \tag{58}$$

with $\underline{S}$, $\underline{p}$, and $\underline{\gamma}$ real and $\underline{S} = -\underline{S}'$.

Solution : Let $\underline{K}$ be given by equation (55). Then since $\underline{K}$ is positive definite it can be written as

$$\underline{P}'\underline{P} = \underline{K} \tag{59}$$

with $\underline{P}$ real and nonsingular. Let $\underline{P}\underline{x} = \underline{z}$ that

$$\dot{\underline{z}} = \underline{PAP}^{-1}\underline{z} + \underline{Pbu} \quad ; \quad y = \underline{c}'\underline{P}^{-1}\underline{z} \tag{60}$$

Now clearly

$$\frac{d}{dt}\,\underline{z}'\underline{z} = \underline{z}'\{\underline{PAP}^{-1} + [\underline{PAP}^{-1}]'\}z + \underline{z}'\underline{Pbu} + u\underline{b}'\underline{Pz} \tag{61}$$

but from equation (55) we see that the $\underline{z}'\{\underline{PAP}^{-1} + [\underline{PAP}^{-1}]'\}z$ is a perfect square and that $\underline{b}'\underline{Pz} = y$. Thus we let

$$2\underline{pp}' = \underline{PAP}^{-1} + [\underline{PAP}^{-1}]' \tag{62}$$

and let

$$\underline{\gamma} = \underline{Pb}$$

then

$$\underline{S} = \underline{PAP}^{-1} - [\underline{PAP}^{-1}]' = -\underline{S}' \tag{63}$$

and all requirements are met. Equation (58) is like the realization shown in figure 2.

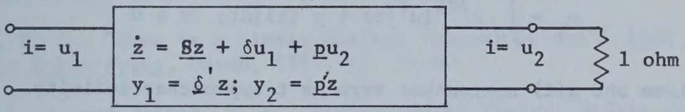

i= u_1
$\dot{\underline{z}} = \underline{S}z + \underline{\delta}u_1 + \underline{p}u_2$
$y_1 = \underline{\delta}'\underline{z}$; $y_2 = \underline{p}'\underline{z}$
i= u_2
1 ohm

Figure 2 : Showing a network equivalent of equation (58)

EXPONENTIALLY WEIGHTED DIFFERENTIALS

Using the operational rule

$$e^{\sigma t}p(D)x = p(D-\sigma)e^{\sigma t}x \tag{64}$$

We can extend our results in various ways. We need the following corollary of lemma 1. (Compare with the special case in Forys [6])

Corollary : There exists an η such that

$$e^{2\sigma t}\sum_{i=0}^{n}\sum_{j=0}^{n}\alpha_{ij}\phi^{(i)}(t)\phi^{(j)}(t) = d\eta/dt \tag{65}$$

if and only if the polynomial

$$h(s) = \sum_{i=0}^{n}\sum_{j=0}^{n}[\alpha_{ij}(s-\sigma)^{i}(-s-\sigma)^{j} + (-s-\sigma)^{i}(s-\sigma)^{j}] \tag{66}$$

vanishes identically.

In this case the expression for the integral will be of the form

$$\int_a^b e^{2\sigma t}\sum_{i=0}^{n}\sum_{j=0}^{n}\alpha_{ij}\phi^{(i)}(t)\phi^{(j)}(t)dt = e^{2\sigma t}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{ij}\phi^{(i-1)}(t)\phi^{(j-1)}(t)\Big|_a^b \tag{67}$$

With this definition of $\underline{K} = \underline{K}'$ we write

$$e^{2\sigma t}\underline{x}'\underline{Kx} = \int_{t(\underline{0})}^{t(\underline{x})} e^{2\sigma t}\sum_{i=0}^{n}\sum_{j=0}^{n}\alpha_{ij}\phi^{(i)}(t)\phi^{(j)}(t)dt \tag{68}$$

It is easy to see that

$$\frac{d}{dt}\underline{x}'\underline{Kx} = -2\sigma\underline{x}'\underline{Kx} + \sum_{i=0}^{n}\sum_{j=0}^{n}\alpha_{ij}x_i x_j \tag{69}$$

These formulae can be used to good advantage in certain problems arising in stability theory [7], however here we use them to derive the solution of a least squares problem.

Problem 7: Given equations (2) and (3) with the state both controllable and observable, find u such that

$$\eta_4 = \int_0^\infty e^{2\sigma t}[u^2(t) + y^2(t)]dt; \quad \sigma \geq 0 \tag{70}$$

is a minimum and $\underline{x}(t)$ approaches zero as t approaches infinity.

Solution: Write η_4 as

$$\eta_4 = \int_0^\infty e^{2\sigma t}[p(D)x_1(t)]^2 + e^{2\sigma t}[q(D)x_1(t)]^2 dt \tag{71}$$

Use the shift theorem to get

$$\eta_4 = \int_0^\infty [p(D-\sigma)x_1(t)e^{\sigma t}]^2 + [q(D-\sigma)x_1(t)e^{\sigma t}]^2 dt \tag{72}$$

add and subtract the quantity $[r(D)x(t)e^{\sigma t}]^2$ where

$$r(D) = [p(D-\sigma)p(-D-\sigma) + q(D-\sigma)q(-D-\sigma)]^+ \tag{73}$$

This gives

$$\begin{aligned}\eta_4 = \int_0^\infty [p(D-\sigma)x_1(t)e^{\sigma t}]^2 + [q(D-\sigma)x_1(t)e^{\sigma t}]^2 - \\ [r(D)x_1(t)e^{\sigma t}]^2 dt + \int_0^\infty [r(D)x_1(t)e^{\sigma t}]^2 dt\end{aligned} \tag{74}$$

From our corollary we see that the values of the first integral is independent of our choice of u(t). We can do no better than to make the second integrand vanish identically. This gives as an equation for $x_1(t)$

$$r(D+\sigma)x_1(t) = 0 \tag{75}$$

and as an equation for u(t)

$$u(t) = p(D)x_1(t) - r(D+\sigma)x_1(t) \tag{76}$$

REFERENCES

1. Carathedory, C., Calculus of Variations and Partial Differential Equations of the First Order, Part 1, Holden Day, San Francisco, 1965.
2. Brockett, R. W., "On the Stability of Nonlinear Feedback Systems," IEEE Trans. Applications and Industry, Nov., 1964, pp. 443-449.
3. Wonham, W. M. and Johnson, C. D., "Optimal Bang-Bang Control with Quadratic Performance Index," ASME Journal of Basic Engineering, March, 1964, pp. 107-115.
4. Kalman, R. E., "When is a Linear Control System Optimal?" ASME Journal of Basic Engineering, March, 1964, pp. 51-60.
5. Zadeh, L. A. and Desoer, C. A., Linear System Theory, McGraw-Hill, 1963.
6. Forys, L. J., Stability of Linear Time-Varying Systems, S.M. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, May, 1965.
7. Gruber, M. and Willems, J.L., "On a generalization of the Circle Criterion," 1966 Allerton Conference on Circuit and System Theory, University of Illinois, Urbana Illinois.

THE ROTATOR--A NEW LINEAR ACTIVE 2-PORT NETWORK ELEMENT

LEON O. CHUA
Purdue University
W. Lafayette, Indiana

ABSTRACT

A new linear, active, reciprocal and bilateral 2-port network element called the Rotator is defined and shown to possess the property that whenever a nonlinear resistor is connected across one of its ports, the characteristic curve seen across the other port is simply the resistor's I-V curve rotated by some prescribed angle.

The Rotator can be realized in the form of a π-network or T-network configuration with two positive and one negative resistors. It can also be realized in the form of a balanced lattice network.

The subjects of stability, power rating and sensitivity are considered and practical stability criterion are derived. Experimental models of the Rotator are reported and experimental data are shown to agree with theoretical predictions. Some potential applications of the Rotator are given.

I. INTRODUCTION

We define the ROTATOR to be a linear, 2-port network element characterized by

$$V_1 = (\mathrm{COS}\ \theta)\, V_2 + (\mathrm{SIN}\ \theta)\, I_2 \qquad (1)$$

$$I_1 = (\mathrm{SIN}\ \theta)\, V_2 - (\mathrm{COS}\ \theta)\, I_2 \qquad (2)$$

The Rotator is therefore completely specified by the parameter θ (in degrees), henceforth called the ANGLE OF ROTATION.

The proposed symbol for the Rotator and its associated 2-port parameters are shown in Table 1.

From Eqs. (1) and (2), we can readily derive the following properties of the Rotator [1]:

Property 1:

If a 2-terminal nonlinear resistor characterized by an I-V curve Γ is connected across port 2 of the Rotator, the resulting I_1-V_1 curve seen across port 1 of the composite network is simply the original curve Γ rotated by θ° (counterclockwise if $\theta > 0$ and clockwise if $\theta < 0$) about the origin.

Property 2:

The Rotator is a bilateral 2-port network element, i.e., property 1 is valid regardless of whether the nonlinear resistor is connected across port 1 or port 2.

Property 3:

The Rotator is a linear, active, reciprocal, frequency-invariant 2-port network element.

Property 4:

The determinant of the open-circuit impedance matrix, the short-circuit admittance matrix, the transmission (chain) matrix or the scattering matrix of the Rotator is either +1 or -1. In particular, the transmission matrix is an orthogonal matrix.

Property 5:

If a load impedance $Z_L(p)$ or a load admittance $Y_L(p)$ is connected across one port of the Rotator, then the driving point impedance $Z_i(p)$ or admittance $Y_i(p)$ across the other port of the Rotator is related to $Z_L(p)$ or $Y_L(p)$ by the following bilinear transformation:

$$Z_i(p) = \frac{(\text{COT}\ \theta)\ Z_L(p) - 1}{Z_L(p) + \text{COT}\ \theta} \qquad (3)$$

$$Y_i(p) = \frac{1 + (\text{COT}\ \theta)\ Y_L(p)}{\text{COT}\ \theta - Y_L(p)} \qquad (4)$$

In particular, when $\theta = 45^\circ$, Eqs. (3) and (4) reduce to the following interesting form:

$$Z_i(p) = \frac{Z_L(p) - 1}{Z_L(p) + 1} \qquad (5)$$

$$Y_i(p) = \frac{1 + Y_L(p)}{1 - Y_L(p)} \qquad (6)$$

Property 6:

If we connect n Rotators (with angle of rotation $\theta_1, \theta_2, \ldots \theta_n$) in cascade, the resulting 2-port network is equivalent to a Rotator with angle of rotation θ, where

$$\theta = \sum_{i=1}^{n} \theta_i \qquad (7)$$

Property 7:

A θ-degree Rotator can be transformed into a ($\theta \pm 180$) - degree Rotator by interchanging the two terminals across either port 1 or port 2 of the Rotator.

Property 8:

The set of all Rotators forms an "Abelian group" [2] with the product operation between two Rotators defined by their cascade connection.

Property 9:

There exist a "physical" as well as mathematical one-to-one correspondence between each point of the rotated curve and that of the corresponding point in the original curve. That is, if point p of the I-V curve of a given resistor becomes point p' after being rotated by θ-degrees by a rotator, and if we operate at point p' of the resulting element, then the given resistor will be operating at point p, and vice-versa.

II. SYNTHESIS OF ROTATORS

A. Networks For Realizing Rotators

Two network configurations for realizing the Rotator with any prescribed angle of rotation are shown in Table 1. The two degenerate cases corresponding to $\theta = 0^\circ$ and $\theta = \pm 180^\circ$ are excluded because the former can be trivially realized by a pair of parallel wires and the later can be realized by a pair of crossed wires. If we choose the network configuration corresponding to the range of θ specified in Table 1, then only one negative resistor is required. Observe that in view of property 7, it is possible to use only one negative resistor and only one network configuration, provided we interchange the terminals across either port 1 or port 2 of the Rotator, whenever θ falls outside of the desired range in Table 1. Since the Rotator is an active 2 - port network, it is impossible to get rid of the negative resistor (except in the trivial case when $\theta = 0^\circ$ and $\theta = \pm 180^\circ$). Hence, the symmetrical π - network or T - network shown in Table 1 is the simplest possible realization of the Rotator in grounded form. If a balanced realization is desired, the balanced symmetrical lattice network shown in Fig. 1 can be used as an alternate realization. Observe that corresponding to any θ, either the series arm elements or the shunt arm elements are absent. The disadvantage of the balanced realization is of course that two negative resistors are invariably required.

Finally, we wish to remark that the element values specified in Table 1 are all referred to an angle of rotation of θ degrees in the ampere-volt plane. If the I-V curve to be rotated is given in the ma-volt plane (which is the case with most practical resistors), it would be necessary to multiply all element values in Table 1 by the corresponding scale factor. Hence, the element values specified in Table 1 can be interpreted as "normalized values". Accordingly, all equations and graphs presented in this paper are also in normalized form.

B. Networks For Realizing Negative Resistance

Since one negative resistance is invariably required in the design of a Rotator, this section is a short survey of various networks for realizing such elements.

The simplest method to realize a negative resistance is to connect a positive resistance across one port of a "Negative Impedance Converter" (NIC). The theory of Negative Impedance Converters is well known and many practical networks using either vacuum tubes or transistors have been proposed and built in the past [3 - 10]. It is also well known that an NIC may be either "Short - Circuit Stable" (SCS) or "Open - Circuit Stable" (OCS); in particular, a "Shunt Type" NIC is short - circuit Stable and a "Series type" NIC is open - circuit Stable [3 - 8]. It will be shown in the next section that a proper choice of an (OCS)-NIC or (SCS)-NIC is crucial to the design of a stable Rotator.

There are other methods to realize negative elements other than using an NIC. For example, any device or interconnection of devices which exhibits a negative resistance characteristic across some driving point terminals can be biased to serve as a negative resistance [11 - 12].

C. Stability Criterion For Designing A Rotator

Since the Rotator is an active 2 - port network element, it is potentially unstable. Hence, before a practical Rotator can be designed, it is necessary that we derive some practical Stability Criterion.

Consider the Rotator networks shown in Figs. 2(a) and 2(b) with load resistance R_L and source resistance R_s. Since the negative resistance is the source of instability, it is necessary for us to determine which type of NIC to use to assure stable operation [1]:

Stability Criterion For The Rotator For a given load resistance R_L and source resistance R_s, the Rotator networks in Figs. 2(a) and 2(b) are stable if, and only if, an (OCS) - NIC is used to realize the negative resistance whenever

$$R_s > \frac{1 - R_L \cot\theta}{R_L + \cot\theta} \tag{8}$$

and a (SCS) - NIC is used whenever

$$R_s < \frac{1 - R_L \cot\theta}{R_L + \cot\theta} \tag{9}$$

The equation

$$R_s = \frac{1 - R_L \cot\theta}{R_L + \cot\theta} \tag{10}$$

is plotted in Fig. 3(a) with the parameter $\theta = 30^\circ$ (-150°), 45° (-135°), 60°(-120°) and 90°(-90°) and in Fig. 3(b) with the parameter $\theta=90^\circ(-90^\circ)$, $120^\circ(-60^\circ)$, $135^\circ(-45^\circ)$ and $150^\circ(-30^\circ)$. Observe that the curves are symmetrical above a 45° line through the origin, and hence the axes for R_s and R_L

may be interchanged at will. In view of the Stability Criterion above, it is clear that for a given value of R_L, R_s and θ, an (OSC) - NIC must be used whenever the point (R_L, R_s) lies above the stability curve corresponding to the specified θ. Otherwise, a (SCS) - NIC must be used to assure stability. Hence the stability curves in Figs. 3(a) and 3(b) reduce the stability problem to a trivial one. In practice, the load is of course invariably a nonlinear resistor whose DC - resistance ranges in general between a lower and upper limits, in which case, the stability curves can still be used to determine the type of NIC to use.

D. Power Dissipation Consideration

If we define $P_s = V_1 I_1$, and $P_L = - V_2 I_2 = VI$, we obtain

$$P_s = \frac{1}{2} (\mathrm{SIN}\ 2\theta)\ (V^2 - I^2) + (\mathrm{COS}\ 2\theta)VI \qquad (11)$$

$$P_L = \frac{1}{2} (\mathrm{SIN}\ 2\theta)\ (I_1^2 - V_1^2) + (\mathrm{COS}\ 2\theta)V_1 I_1 \qquad (12)$$

Eqs. (11) and (12) are useful in relating the power dissipated in the load to the input power and vice-versa. If the load resistor has a maximum power rating $P_L < P_{L_{max}}$, the rotated element must be restricted to

$$\frac{1}{2}(\mathrm{SIN}\ 2\theta)(I_1^2 - V_1^2) + (\mathrm{COS}\ 2\theta)V_1 I_1 < P_{L_{max}} \qquad (13)$$

Observe that Eq. (13) represents a rotated hyperbola and the maximum power curve in the I_1-V_1 plane as defined by Eq. (13) is simply the maximum power hyperbola in the I-V plane rotated by θ - degrees. The above interpretation provides us a simple means to determine the power ratings of the rotated element. Conversely, if a certain rotated element is to be synthesized with a given maximum power rating, the minimum power rating of the load resistor can be easily found by the above interpretation.

In order to find the power rating of resistors G_1, G_2 and G_3 of the π - network Rotator, we found

$$P_1 = \mathrm{TAN}\ \frac{\theta}{2}\ [(\mathrm{COS}\ \theta)\ V - (\mathrm{SIN}\ \theta)\ I]^2, P_2 = (\mathrm{TAN}\ \frac{\theta}{2})V_2, P_3 = -[I + (\mathrm{TAN}\ \frac{\theta}{2})V]^2 \mathrm{SIN}\ \theta \qquad (14)$$

Similarly, the power rating of resistors R_1, R_2 and R_3 of the T - network Rotator are given respectively by

$$P_1' = -\ \mathrm{TAN}\ \frac{\theta}{2}[(\mathrm{SIN}\ \theta)V + (\mathrm{COS}\ \theta)I]^2, P_2' = -(\mathrm{TAN}\ \frac{\theta}{2})I^2, P_3' = \mathrm{SIN}\ \theta[V - (\mathrm{TAN}\ \frac{\theta}{2})I]^2 \qquad (15)$$

Observe that Eqs. (14) and (15) represent a family of straight lines in the I-V plane with the respective power P_1, P_2, P_3, P_1', P_2' or P_3' as a parameter. By superimposing these "constant power straight lines" on top of the I-V plane, the maximum power rating of each resistor of the Rotator corresponding to any nonlinear load resistor can be easily determined.

E. Sensitivity Consideration

Since very stable and accurate positive resistances are readily available in practice, we would be concerned only with the sensitivity of the Rotator to variations in the value of the negative resistance G_3 for the π - network Rotator and R_3 for the T - network Rotator. Since the main purpose of the Rotator is to rotate a given I-V curve, any meaningful definition of Rotator Sensitivity must be a measure of the departure or "distortion" of the desired rotated curve from the actual curve. The following two definitions are proposed to give a measure of the Rotator Sensitivity:

Let P with coordinates (I_1, V_1) be a point in the I_1-V_1 plane representing the true location of a point on a rotated I-V curve. Let P$'$ with coordinates (I_1', V_1') be the point in the I_1-V_1 plane as a result of variation in the desired values of G_3 or R_3. If we let $\rho\angle\theta$ and $\rho'\angle\theta'$ be the polar coordinates of P and P$'$ respectively as shown in Fig. 4(a), and let d be the distance between point P and P$'$, then we can define the following:

Definition 1: Relative Rotator Error

We will define the "Relative Rotator Error" ϵ_R to be the ratio between d and ρ, i.e.

$$\epsilon_R \equiv \frac{d}{\rho} = \frac{\sqrt{(V_1'-V_1)^2 + (I_1'-I_1)^2}}{\sqrt{I_1^2 + V_1^2}} \tag{16}$$

Definition 2: Rotator Phase and Magnitude Error

We will define the "Rotator Phase Error ϵ''_θ and "Rotator Magnitude Error ϵ''_ρ respectively by

$$\epsilon_\theta \equiv \theta' - \theta = \mathrm{TAN}^{-1} \frac{I_1'}{V_1'} - \mathrm{TAN}^{-1} \frac{I_1}{V_1} \tag{17}$$

$$\epsilon_\rho \equiv \rho' - \rho = \sqrt{V_1'^2 + I_1'^2} - \sqrt{V_1^2 + I_1^2} \tag{18}$$

Eq. (17) can be expressed into the more convenient form

$$\epsilon_\theta = \mathrm{TAN}^{-1} \left[\frac{I_1'V_1 - I_1V_1'}{V_1'V_1 + I_1'I_1} \right] \tag{19}$$

A closer study of the above definitions shows that Definition 1 is a more meaningful measure when the entire I-V curve is of interest. Definition 2 however will be more meaningful if specific points or small regions of the I-V curve is of interest. If we let

$$\epsilon \equiv \frac{\Delta G_3}{G_3} \equiv \frac{\Delta R_3}{R_3} \tag{20}$$

we can readily derive the following useful expressions:

$$\epsilon_R = \frac{2\epsilon}{(1+\epsilon)} \left\{ \frac{\mathrm{SIN}\,\frac{\theta}{2}\,[(\mathrm{TAN}\,\frac{\theta}{2})\,V + I]}{\sqrt{V^2 + I^2}} \right\} \tag{21}$$

$$\epsilon_\theta = \mathrm{TAN}^{-1} \left\{ \frac{\mathrm{CSC}\,\theta\,[(\mathrm{COS}\,\theta - 1)\,V - (\mathrm{SIN}\,\theta)\,I]^2}{[\mathrm{COS}\,\theta - 1][V^2 + (2\,\mathrm{COT}\,\theta)\,VI - I^2] - (\frac{1+\epsilon}{\epsilon})(V^2+I^2)} \right\} \tag{22}$$

$$\epsilon_\rho = \frac{1}{(1+\epsilon)} \sqrt{(V^2+I^2) + (\epsilon^2 A + 2\epsilon B)} - \sqrt{V^2 + I^2} \tag{23}$$

where

$$A \equiv \left\{ \frac{(5 - 8\,\mathrm{COS}\,\theta + 3\,\mathrm{COS}^2\,\theta)}{\mathrm{SIN}^2\,\theta} \right\} V^2 + I^2 + \left[\frac{4\,(1 - \mathrm{COS}\,\theta)}{\mathrm{SIN}\,\theta} \right] VI \tag{24}$$

$$B \equiv \left\{ \frac{2 - \mathrm{COS}\,\theta - 3\,\mathrm{COS}^2\theta + 2\,\mathrm{COS}^3\theta}{\mathrm{SIN}^2\theta} \right\} V^2 + (\mathrm{COS}\,\theta)\,I^2 + \left[\frac{2\,\mathrm{COS}\,\theta\,(\mathrm{COS}\,\theta - 1)}{\mathrm{SIN}\,\theta} \right] VI \tag{25}$$

Observe that the above Rotator errors are nonlinear functions of I and V. Hence the sensitivity of the Rotated curves is different on different points of the I-V curve. To obtain a conservative estimate on the Rotator sensitivity we derive for the maximum value of ϵ_R in terms of ϵ and θ directly:

$$\epsilon_{R_{max}} = \left[\frac{2\epsilon}{1+\epsilon} \right] \mathrm{TAN}\,\frac{\theta}{2} \tag{26}$$

Observe that Eq. (26) is independent of the I-V curve of the element to be rotated and represents an "upper bound" for the Relative Rotator error ϵ_R. Eq. (26) is plotted in Fig. 4(b) as a function of ϵ with θ as a parameter. Observe that the Maximum Relative Rotator Error increases as θ increases.

Hence, a Rotator with a larger θ is more sensitive to variation in the values of G_3 or R_3 than a Rotator with a smaller θ. Observe also that for $\theta < 50^\circ$, the percentage Relative Error is less than the percentage variation of G_3 or R_3. For example, a 20% change in G_3 or R_3 of a 10° - Rotator results in a 2.8% "Maximum" Relative Error. Hence, qualitatively, we can say that for small angles of rotation, the Rotator is rather insensitive to variations in G_3 or R_3.

III. EXPERIMENTAL RESULTS

In order to verify the theory presented above, various laboratory models of the π-network configuration Rotator and the T-network configuration Rotator have been built. The negative resistance in the earlier models makes use of the operational amplifier NIC circuit shown in Fig. 5.

Some of the experimental results are shown in Figs. 6 to 9. Fig. 6(a) shows the I-V curve of a 2-terminal nonlinear resistor being rotated by 30°, 45°, 60°, 90° and 135° respectively in the counterclockwise direction. Fig. 6(b) shows the same I-V curve being rotated by -30°, -45°, -60°, -90°, -135° and -150° degrees respectively in the clockwise direction. Fig. 7 shows the collector-to-emitter characteristic of a PNP Transistor being rotated by 2°, 30° in the counterclockwise direction and -5°, -10° and -30° in the clockwise direction. The "loops" that appear in some of these pictures are not inherent in the Rotator, but instead come from the external "Tracing Circuit" and the grounding problem associated with the operational amplifier and the oscilloscope.

The stability criterion derived in sec. II-C was verified experimentally by the data shown in Figs. 8(a) and 8(b). The solid curves shown in these figures represent the theoretical values as predicted in Figs. 3(a) and 3(b). Observe that the agreement between the experimental data and that of the predicted data is quite good. The discrepancy from the predicted value can be accounted for by the tolerance of the resistors, the low quality operational amplifier used in the experiments and the inaccuracies in the measurement process.

The sensitivity of the Rotator to variations in the values of G_3 and R_3 are shown by the multiple-exposure pictures in Fig. 9. Observe that there are four traces in each picture. The uppermost trace corresponds to the case when the value of G_3 or R_3 is exact. The second trace corresponds to the case where the value of G_3 or R_3 is changed by 10%. The third trace corresponds to the case where the value of G_3 or R_3 is changed by 20%. The lowermost trace corresponds to the case where the value of G_3 or R_3 is changed by 50%. Observe that as the angle of rotation θ increases, the Rotator becomes more sensitive in the sense that the rotated curve suffers more distortion from the original curve. This fact agrees very well with the predictions in Sec. II-E.

IV. SOME POTENTIAL APPLICATIONS OF THE ROTATOR

Aside from making available an entirely new generation of nonlinear circuit elements with characteristic curves unobtainable before, a Rotator can be used to modify device parameters such as gain, input impedance, output impedance of existing devices such as Transistor, vacuum tubes, etc., and thus improve, if not optimize the performance of the associated circuitry. As a result of the availability of new characteristic curves, new and novel functional circuits may be designed to take advantage of the new characteristics. For example, with a suitable choice of nonlinear element and a Rotator, a more accurate square law device may be realized.

The Rotator can be used as a basic component to realize resistors with prescribed non-monotonic I-V curves. The Rotator can also be used to trace a class of multi-valued I-V curves in an oscilloscope, a task that is impossible with present techniques.

A Rotator can be used as an element for active network synthesis. For angles of rotation $\theta < 45^\circ$, the Rotator as an active element is, qualitatively speaking, less sensitive than an NIC and hence might be preferred over other existing active elements.

Conceptually, the Rotator can be used as a basic component for synthesizing "models" of 3 - terminal elements. By combining the elements of the Rotator with other elements in the model, the Rotator can be absorbed. In this respect, the Rotator need never be built, it is being used only as a gimmick to synthesize nonlinear models.

The above are just a few of the many potential applications of the Rotator. It is hoped that more applications will be conceived of in the future.

V. CONCLUDING REMARKS

We have shown that the Rotator is a simple and inexpensive electronic device. A stable Rotator terminated in a prescribed nonlinear load resistor and source resistance can be easily designed using the stability criterion formulated in Sec. II-C. The power rating of the rotated element can be determined from the rotated maximum power hyperbola of the nonlinear load resistor. The Rotator is seen to be relatively insensitive to parameter variations for $\theta<90^\circ$ and becomes quite sensitive for $90^\circ<\theta<180^\circ$. However, in view of properties 6 and 7, it is possible to minimize the sensitivity for large angles by either cascading two or more Rotators with smaller angles of rotation, or by interchanging a pair of terminals to obtain a 180° rotation about the origin.

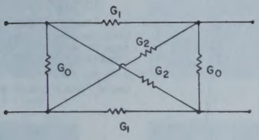
Several models of Rotators have been built using various different NIC circuits. The latest model has a dimension $\frac{1}{2}$" x 2" x 4" and uses 6 transistors. Effort is presently going on to integrate the entire circuit into the size of a transistor. As a practical component, the Rotator is completely specified by its angle of rotation and its power handling capability.

REFERENCES

1. Chua, L.O., "The Rotator - A New Electronic Device", Semi-Annual Research Summary, School of Elect. Eng., Purdue University, Laf., Ind., June, 1966.
2. Birkhoff, G. and S. MacLane, A Survey of Modern Algebra, Macmillan Company, 1953.
3. Crisson, G., "Negative Impedances and the Twin 21-type Repeater", Bell Sys. Tech. Jour., Vol. 10, pp. 485-513; July, 1931.
4. Merill, J.L., Jr., "Theory of the Negative Impedance Converter", BSTJ, Vol. 30, Jan., 1951, pp. 88-109.
5. Ginzton, E.L., "Stabilized Negative Impedances", Electronics, Vol. 18, (July, Aug., Sept., 1945), pp. 140-150, 138-145, 140-144.
6. Linvill, J.G., "Transistor Negative Impedance Converters", Proc. of IRE, Vol. 41, June, 1953, pp. 725-729.
7. Larky. A.I., "Negative Impedance Converters", IRE Trans. Circuit Theory, Vol. CT-4, Sept., 1957, pp. 124-131.
8. Yanagisawa, T., "Current Inversion Type Negative Impedance Converters", J. Inst. Elec. Commun. Engrs. (Japan), Vol. 39, pp. 933-937, Nov., 1956.
9. Lundry, W.R., "Negative Impedance Circuits: Some Basic Relations and Limitations", IRE Trans. Circuit Theory", Vol. CT-4, Sept. 1957, pp. 132-139.
10. Blecher, F.H., "Application of Synthesis Techniques to Electronic Circuit Design", IRE Trans. Circuit Theory, Vol. CT-7, Special Supplement, 1960, pp. 79-91.
11. Herold, E.W., "Negative Resistance and Devices for Obtaining It", PROC. IRE, Vol. 23, pp. 1201-1223, 1935.
12. Su, K.L., Active Network Synthesis, McGraw-Hill Book Co., 1965, pp. 9-26.

Acknowledgment

This research is supported in part
by NSF Grant GK-22.


G₁
G₂
G₀
G₂
G₀
G₁

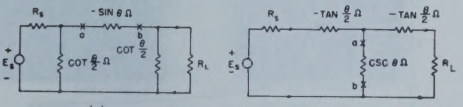
$G_0 = -\cot\theta - |\csc\theta|$

$G_1 = |\csc\theta| - \csc\theta$

$G_2 = |\csc\theta| + \csc\theta$

Fig. 1

A Balanced Realization of Rotator


Rₛ
- SIN θ Ω
a
b
COT θ/2
+
Eₛ
-
COT θ/2 Ω
R_L
Rₛ
-TAN θ/2 Ω
-TAN θ/2 Ω
a
b
CSC θ Ω
R_L
+
Eₛ
-

(a) (b)

Fig. 2

A Rotator With Load Resistance R_L and Source Resistance R_S.

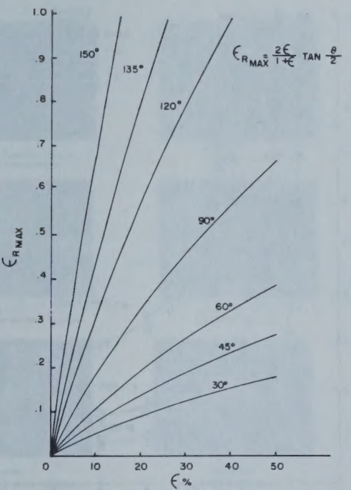


Fig 4 (b)

MAXIMUM RELATIVE ROTATOR ERROR VS PERCENT ERROR IN G_3 OR R_3.

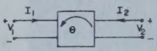
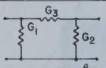
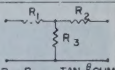
	THE ROTATOR  	
REMARKS: All element values must be multiplied by a scale factor determined by the scale of the curve to be rotated. EXAMPLE: To rotate an I-V curve in the ma-Volt plane, multiply the values of R_1, R_2 and R_3 by 10^3	Π-Configuration  $G_1 = G_2 = \tan\frac{\theta}{2}$ MHO $G_3 = -\csc\theta$ MHO	T-Configuration  $R_1 = R_2 = -\tan\frac{\theta}{2}$ OHM $R_3 = \csc\theta$ OHM
Range of θ Which Requires Only One Negative Element	$0 < \theta < \pi$ (Counter-Clockwise)	$-\pi < \theta < 0$ (Clockwise)
Open Circuit Impedance Matrix $\underset{\sim}{Z}(p)$	$\underset{\sim}{Z}(p) = \begin{bmatrix}\cot\theta & \csc\theta \\ \csc\theta & \cot\theta\end{bmatrix}$	
Short Circuit Admittance Matrix $\underset{\sim}{Y}(p)$	$\underset{\sim}{Y}(p) = \begin{bmatrix}-\cot\theta & \csc\theta \\ \csc\theta & -\cot\theta\end{bmatrix}$	
Transmission (Chain) Matrix $\underset{\sim}{T}(p)$	$\underset{\sim}{T}(p) = \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}$	
Scattering Matrix $\underset{\sim}{S}(p)$ (Normalized to Unit Port Numbers)	$\underset{\sim}{S}(p) = \begin{bmatrix}-\tan\theta & \sec\theta \\ \sec\theta & -\tan\theta\end{bmatrix}$	

TABLE I

TWO PORT PARAMETERS OF ROTATOR

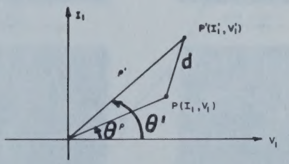

I₁
V₁
P'(I₁', V₁')
P(I₁, V₁)
ρ'
d
θᵖ
θ'

Fig 4 (a)

Symbols Associated With The Definition of Rotator Sensitivity.

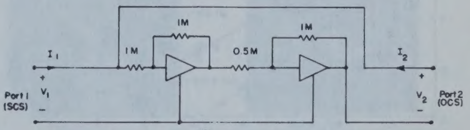

1M
1M
0.5M
1M
I₁
I₂
+
V₁
-
Port 1 (SCS)
+
V₂
-
Port 2 (OCS)

Fig 5

Operational Amplifier Realization of NIC

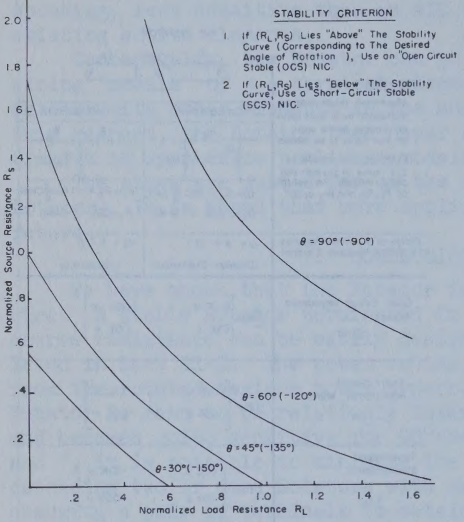


Fig. 3 (a)

STABILITY CRITERION FOR PROPER CHOICE OF NEGATIVE IMPEDANCE CONVERTERS FOR VARIOUS ANGLES OF ROTATION: = 30° (-150°), 45° (-135°), 60° (-120°), 90° (-90°)

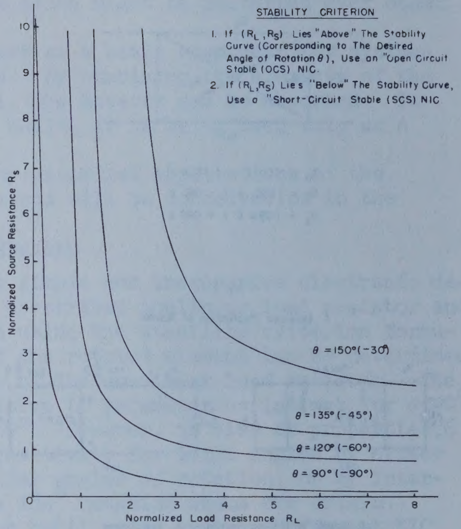


Fig. 3 (b)

STABILITY CRITERION FOR PROPER CHOICE OF NEGATIVE IMPEDANCE CONVERTERS FOR VARIOUS ANGLES OF ROTATION: θ = 90° (- 90°), 120°(- 60°), 135° (- 45°), 150°(-30°)

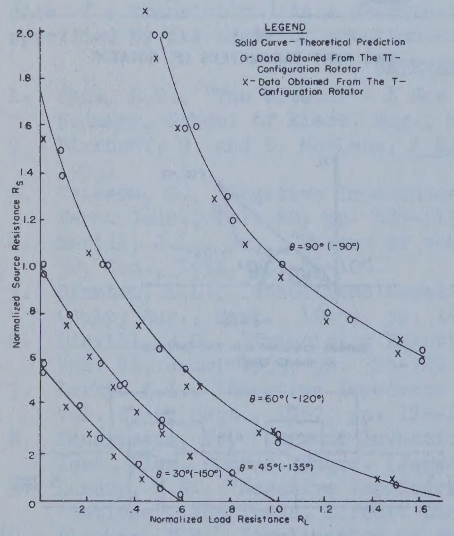


Fig. 8 (a)

EXPERIMENTAL VERIFICATION OF STABILITY CRITERION OF ROTATOR: θ = 30° (-150°), 45° (-135°), 60° (-120°), 90° (-90°)

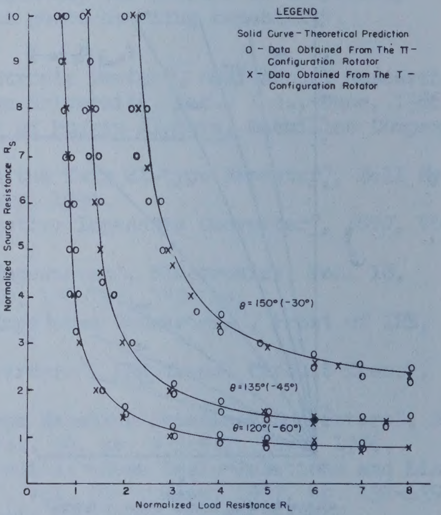
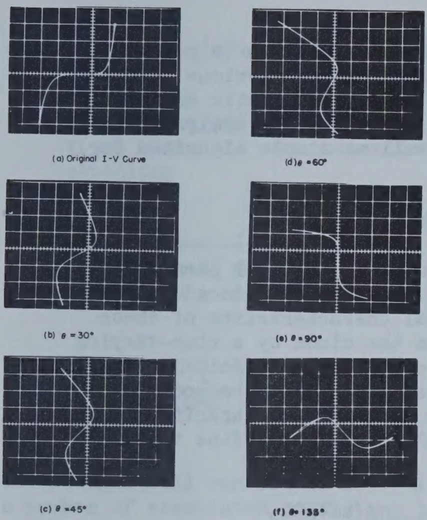


Fig. 8 (b)

EXPERIMENTAL VERIFICATION OF STABILITY CRITERION OF ROTATOR: θ = 120° (-60°), 135° (-45°), 150° (-30°)



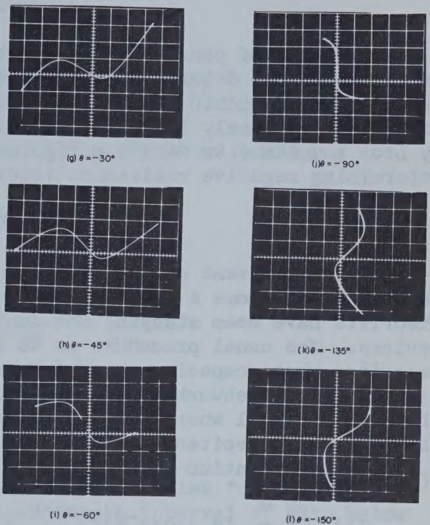
(a) Original I-V Curve (b) θ = 30° (c) θ = 45° (d) θ = 60° (e) θ = 90° (f) θ = 135°

Vertical Axis — I (ma)
Horizontal Axis — V (Volts)

Scale = 0.2 ma/cm
Scale = 0.2 Volt/cm

Figure 6 (a)

Counter-Clockwise Rotation of an I-V Curve by a Rotator Using The Π – Network Confguration.



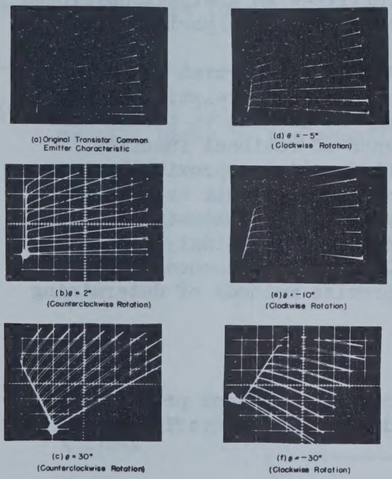
(g) θ = −30° (h) θ = −45° (i) θ = −60° (j) θ = −90° (k) θ = −135° (l) θ = −150°

Vertical Axis – I (ma)
Horizontal Axis – V (Volts)

Scale = 0.2 ma/cm
Scale = 0.2 Volt/cm

Figure 6 (b)

Clockwise Rotation of an I-V Curve by a Rotator Using The T-Network Configuration.



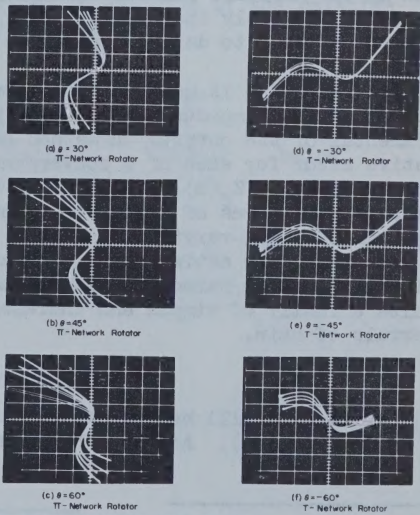
(a) Original Transistor Common Emitter Characteristic
(b) θ = 2° (Counterclockwise Rotation)
(c) θ = 30° (Counterclockwise Rotation)
(d) θ = −5° (Clockwise Rotation)
(e) θ = −10° (Clockwise Rotation)
(f) θ = −30° (Clockwise Rotation)

Vertical Axis – Collector Current in ma
Horizontal Axis – Collector-to-Emitter Voltage in Volts
Base Current Steps – 5 μa/step

Scale = 1 ma/cm
Scale = 1 Volt/cm

Figure 7

ROTATION OF THE COLLECTOR-TO-EMITTER CHARACTERISTIC OF A PNP TRANSISTOR (2N404A) BY A ROTATOR.



(a) θ = 30° Π-Network Rotator
(b) θ = 45° Π-Network Rotator
(c) θ = 60° Π-Network Rotator
(d) θ = −30° T-Network Rotator
(e) θ = −45° T-Network Rotator
(f) θ = −60° T-Network Rotator

Remarks: The Upper Most Curve Corresponds to ε = 0%
The Second Curve Corresponds to ε = 10%
The Third Curve Corresponds to ε = 20%
The Lower Most Curve Corresponds to ε = 50%

Figure 9

SENSITIVITY OF R-ROTATOR TO VARIATIONS IN THE VALUES G_3 OR R_3

A NEW METHOD FOR THE DETERMINATION OF GAIN IN PARAMETRIC AMPLIFIERS

*J. V. ADAMS and D. R. ANDERSON
Purdue University
Lafayette, Indiana

ABSTRACT**

A problem of considerable importance in the design of a parametric amplifier is the determination of intervals of gain. Previous results based on the Harmonic Balance analysis are overly optimistic and are in some cases extremely inaccurate. This paper utilizes an analysis developed by Leon and Adams to derive a rigorous as well as simple algorithm for determining negative resistance intervals.

1. INTRODUCTION

With the advent of the semiconductor diode, practical parametric devices have become a reality. For a little more than a decade, network theorists have been studying the fundamental characteristics of these devices. The usual procedure is to replace the diode by a time-varying capacitor whose capacitance as a function of time is sinusoidal. A linear time-invariant network containing such a capacitance may be modeled as shown in Figure 1 where $Y(\omega)$ is assumed to have a shunt capacitance $C_0>2$ so that the net capacitance is positive for all time. We define the impedance $Z_0(\omega)$ by the relation

$$v(t) = \sum_{k=-\infty}^{\infty} [Z_k(\omega) I e^{-j(\omega_s+k)t} + Z_k(\omega)^* I^* e^{j(\omega_s+k)t}]$$

A problem of considerable importance in the design of a parametric amplifier is the determination of intervals of gain. Previous results based on the Harmonic Balance analysis are overly optimistic and are in some cases extremely inaccurate. This paper utilizes an analysis developed by Leon and Adams to derive a rigorous as well as simple algorithm for determining.

The algorithm is motivated by Leon and Adam's [1] partial fraction expansion of the impedance $Z_0(\omega)$ of a parametric circuit relative to the fundamental of the output. A method of Perron [2] is used to obtain the relative error for each of a convergent sequence of rational function approximations to $Z_0(\omega)$. The various rational function approximations are expressible in terms of $Y(j\omega)$, the admittance of the circuit seen looking back from the time-varying capacitor. Moreover, the approximations, together with the error estimates, yields a sequence of increasingly accurate refinements of the Harmonic Balance analysis. As a consequence, there results a family of simple and increasingly precise methods of determining intervals of gain.

2. RESULTS

Our results will be specifically concerned with various properties of the impedance $Z_0(\omega)$. As shown by Leon and Adams, $Z_0(\omega)$ satisfies the following formula for real ω (the frequency of the capacitor variation being normalized to one):

---

*J. V. Adams is currently with Autonetics, Anaheim, California

**This research was supported by the National Science Foundation, Grant GK26.

$$Z_0(\omega) = [j\omega[L(\omega) + A(\omega) + L^*(-\omega)]]^{-1} \tag{2.1}$$

where $L(\omega)$ and $A(\omega)$ are given below

$$j\omega\, A(\omega) = Y(j^{\omega}) \tag{2.2}$$

$Y(j\omega)$ being the admittance looking back from the time-varying capacitor

$$L(\omega) = \cfrac{-1}{A(\omega-1) + \cfrac{(-1)}{A(\omega-2) + \cdots + \cfrac{}{A(\omega-n)} + \cdots}} \tag{2.3}$$

$$\triangleq \left(\frac{-1}{A(\omega-n)}\right)_{n=1}^{\infty} \; (*)$$

Our results will consist of several approximation formulas for $Z_0(\omega)$ and a number of associated sufficient conditions for an interval of gain below $\omega = 1$, i.e., for $\mathrm{Re}[Z_0(\omega)]$ to be negative when ω is in some interval below $\omega = 1$. One such sufficient condition turns out to be a criterion if $Y(j\omega)$ is a sufficiently close uniform approximation to a capacitive admittance for $\omega \geqq 1$. Because of the limited space available, proofs for the various theorems are not given.

We present first a theorem which gives an estimate of the relative error introduced into $\mathrm{Re}[Z_0(\omega)]$ if $L(\omega)$ is truncated.

(*) Throughout this paper the symbols $\left(\frac{a_n}{b_n}\right)_{n=1}^{N}$ and $\left(\frac{a_n}{b_n}\right)_{n=1}^{\infty}$ shall have the meaning given by the equations below

$$\left(\frac{a_n}{b_n}\right)_{n=1}^{N} = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cdots + \cfrac{a_N}{b_N}}} \tag{2.3a}$$

$$\left(\frac{a_n}{b_n}\right)_{n=1}^{\infty} = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cdots + \cfrac{a_n}{b_n + \cdots}}} \tag{2.3b}$$

Theorem 1

Suppose that for some $C_0 > 2$ and some ω such that $0 < \omega < 1$ we have

$$|A(\pm\omega-n) - C_0| < \varepsilon < \frac{C_0}{2} \; ; \; |\omega-n| > 1 \tag{2.4}$$

and

$$\varepsilon < \frac{\rho_1 (1-\theta)^2}{2 + 8\theta - 2\theta^2} \tag{2.5}$$

where

$$\rho_1 = \frac{C_0 + \sqrt{C_0^2-4}}{2}$$

$$\theta = \frac{\quad}{[C_0 + \sqrt{C_0^2-4}}$$

Then we have the following estimate of the relative error made in using

$$\mathrm{Re}[Z_0^{(n)}(\omega)] = \mathrm{Re}[[j\omega(L_n(\omega) + A(\omega) + L_n^*(-\omega))]^{-1}]$$

in the place of $\mathrm{Re}[Z_0(\omega)]$ where $L_n(\omega) = \left[\frac{(-1)}{A(\omega-k)}\right]_{k=1}^{n}$

$$\mathrm{Re}[Z_0(\omega) - Z_0^{(n)}(\omega)] / \mathrm{Re}[Z_0(\omega)] \leq (e_n/f_n)[1 + (e_n/f_n)] \tag{2.6}$$

where

$$e_n = \frac{(1-\theta^2)[1 - \theta - (k+1)s](1-s)}{(C_0-\varepsilon-\frac{1}{\rho_1})\,[1 - (1+s)\theta - (2k+1)s](1-\theta^n)} \left(\frac{k}{1-s}\right)^{n-1}$$

$$\begin{aligned} X \Bigg\{ 1 + \Bigg[\Bigg\{ \Bigg| A(\omega-1) + \frac{(-1)}{A(\omega-2) - \frac{1}{\rho_1}} \Bigg| - \Bigg(\frac{1}{C_0-\varepsilon-\frac{1}{\rho_1}} \Bigg) \Bigg| \frac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{[1-(1+s)\theta-(2k+1)s](1-\theta^{n-1})} \left(\frac{k}{1-s}\right)^{n-2} \\ + \frac{(k+s)\theta}{[1-(1+s)\theta-(2k+1)s]} \Bigg]\Bigg\}^{-1} \Bigg\{ \Bigg| A(\omega-1) + \frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}} \Bigg| \\ - \frac{(k+s)\theta}{(C_0-\varepsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]} \Bigg\}^{-1} \end{aligned}$$

$$f_n = \left\{ \begin{array}{l} \left| \mathrm{Im} \left\{ \dfrac{(-1)}{A(\omega-1) + \dfrac{(-1)}{A(\omega-2) - \dfrac{1}{\rho_1}}} + A(\omega) + \dfrac{(-1)}{A^*(-\omega-1) - \dfrac{1}{\rho_1}} \right\} \right| \\ - \left\{ \left(\dfrac{1}{C_0 - \varepsilon - \dfrac{1}{\rho_1}} \right) \left[\dfrac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{[1-(1+s)\theta-(2k+1)s](1-\theta^{n-1})} \left(\dfrac{k}{1-s} \right)^{n-2} + \dfrac{(k+s)\theta}{[1-(1+s)\theta-(2k+1)s]} \right] \right\} \\ \left\{ 1 + \left| A(\omega-1) + \dfrac{(-1)}{A(\omega-2) - \dfrac{1}{\rho_1}} \right|^{-1} \left[\left| A(\omega-1) + \dfrac{(-1)}{A(\omega-2) - \dfrac{1}{\rho_1}} \right| \right. \right. \\ \left. \left. - \dfrac{1}{C_0 - \varepsilon - \dfrac{1}{\rho_1}} \left[\dfrac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{[1-(1+s)\theta-(2k+1)s](1-\theta^{n-1})} \left(\dfrac{k}{1-s} \right)^{n-2} + \dfrac{(k+s)\theta}{[1-(1+s)\theta-(2k+1)s]} \right] \right] \right\}^{-1} \end{array} \right.$$

$$k = 2\theta / [1+\theta-\varepsilon/\rho_1 + \sqrt{(1-\theta+\varepsilon/\rho_1)^2 - 4\varepsilon/\rho_1}],$$

$$s = 2(\varepsilon/\rho_1)/[1-\theta+\varepsilon/\rho_1 + \sqrt{(1-\theta+\varepsilon/\rho_1)^2 - 4\varepsilon/\rho_1}],$$

$$\theta = \frac{4}{[C_0 + \sqrt{C_0^2 - 4}]^2}$$

The next theorem provides a similar estimate for the relative error introduced into $\mathrm{Re}[Z_0(\omega)]$ by assuming $A(\omega) = C_0$ when $|\omega| \geqq 2$.

Theorem 2

With the same hypotheses as in Theorem 1, the relative error in replacing $\mathrm{Re}[Z_0(\omega)]$ by

$$\mathrm{Re} \left\{ j\omega \left(\frac{(-1)}{A(\omega-1) + \dfrac{(-1)}{A(\omega-2) - \dfrac{1}{\rho_1}}} + A(\omega) + \frac{(-1)}{A^*(-\omega-1) - \dfrac{1}{\rho_1}} \right) \right\}^{-1}$$

can be estimated as follows:

$$\frac{\left| \mathrm{Re}\, Z_0(\omega) - \left[j\omega \left(\dfrac{(-1)}{A(\omega-1) + \dfrac{(-1)}{A(\omega-2) - \dfrac{1}{\rho_1}}} + A(\omega) + \dfrac{(-1)}{A^*(-\omega-1) - \dfrac{1}{\rho_1}} \right) \right]^{-1} \right|}{|\mathrm{Re}[Z_0(\omega)]|} \tag{2.7}$$

$$\leqq (e/f)[1+(e/f)]$$

where

$$e = \frac{(k+s)\theta}{(C_0-\epsilon-\frac{1}{\rho_1})\,[1-(1+s)\theta-(2k+1s]} \quad X$$

$$X\left\{1+\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|^{-1}\left[\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|-\frac{(k+s)\theta}{(C_0-\epsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]}\right]^{-1}\right.$$

$$f = \left|\ \mathrm{Im}\left[\frac{(-1)}{A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}}+A(\omega)+\frac{(-1)}{A^*(-\omega-1)-\frac{1}{\rho_1}}\right]\right|,$$

$$k = 2\theta/[1+\theta-\epsilon/\rho_1+\sqrt{(1-\theta+\epsilon/\rho_1)^2-4\epsilon/\rho_1}],$$

$$s = 2(\epsilon/\rho_1)/[1-\theta+\epsilon/\rho_1+\sqrt{(1-\theta+\epsilon/\rho_1)^2-4\epsilon/\rho_1}],$$

$$\theta = \frac{4}{[C_0+\sqrt{C_0{}^2-4}]^2}$$

The next two theorems give the conditions under which $\mathrm{Re}[Z_0(\omega)]$ is negative.

Theorem 3

If we have the same hypotheses as in Theorem 1 and if

$$\left(\frac{1}{C_0-\epsilon-\frac{1}{\rho_1}}\right)\left(\frac{(1-\theta^2)[1-\theta-(1+k)s](1-s)}{[1-(1+s)\theta-(2k+1)s](1-\theta^n)}\left(\frac{k}{1-s}\right)^{n-1}+\frac{(k+s)\theta}{[1-(1+s)\theta-(2k+1)s]}\right)$$

$$< \left|A(\omega-1)+\frac{-1}{A(\omega-2)-\frac{1}{\rho_1}}\right|, \tag{2.8}$$

then

$$\mathrm{Re}[Z_0(\omega)] < 0$$

if and only if there exists an n such that

$$e_n < |\ \mathrm{Im}[L_n(\omega)+L_n{}^*(-\omega)+A(\omega)]\ |$$

where (2.9)

$$e_n = \frac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{(C_0-\epsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s](1-\theta^n)}\left(\frac{k}{1-s}\right)^{n-1} \quad X$$

$$\times\left[1+\left\{\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|-\left(\frac{1}{C_0-\varepsilon-\frac{1}{\rho_1}}\right)\left[\frac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{[1-(1+s)\theta-(2k+1)s](1-\theta^n)}\left(\frac{k}{1-s}\right)^{n-2}+\frac{(k+s)\theta}{[1-(1+s)\theta-(2k+1)s]}\right]\right\}^{-1}\left\{\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|-\frac{(k+s)\theta}{(C_0-\varepsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]}\right\}^{-1}\right.$$

Moreover, under the hypotheses of Theorem 1, $\mathrm{Re}[Z_0(\omega)] < 0$ **if and only if** there exists an n such that

$$e'_n < \mathrm{Im}[L_n(\omega) + A(\omega) + L_n^*(-\omega)] \tag{2.10}$$

$$e'_n = \frac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{(C_0-\varepsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s](1-\theta^n)}\left(\frac{k}{1-s}\right)^{n-1}$$

$$\times\left[1+\left\{|A(\omega-1)+L_{n-1}(\omega-1)|-\frac{(1-\theta^2)[1-\theta-(k+1)s](1-s)}{(C_0-\varepsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s](1-\theta^{n-1})}\left(\frac{k}{1-s}\right)^{n-2}\right\}^{-1}\times|A(\omega-1)+L_{n-1}(\omega-1)|^{-1}\right.$$

Theorem 4

If we assume the hypotheses of Theorem 1, if we have

$$\frac{(k+s)\theta}{(C_0-\epsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]} < \left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right| \tag{2.11}$$

and if we have

$$e < \left|\mathrm{Im}\left(\frac{(-1)}{A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}}+A(\omega)+\frac{(-1)}{A^*(-\omega-1)-\frac{1}{\rho_1}}\right)\right| \tag{2.12}$$

where

$$e = \frac{(k+s)\theta}{(C_0-\varepsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]}\left[1+\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|^{-1}\right]$$

$$\times\left(\left(\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|-\frac{(k+s)\theta}{(C_0-\varepsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]}\right)^{-1}\right)$$

then

$$\mathrm{Re}[Z_0(\omega)] < 0.$$

3. COMMENTS

The most important comment to be made about the estimates occurring in the four theorems above is that they have relatively simple and very accurrate approximate forms in most interesting cases. Current practical design methods cause $|A(\omega)-C_0|$ to be much less that $C_0/2$ once ω is only moderately larger than 1. Thus one usually has (see Figure 2)

$$\epsilon/\rho_1 << 1-\theta$$

That is, one usually has

$$k \simeq \theta$$

$$s \simeq 0$$

$$\frac{(1-\theta^2)[1-\theta-(k+1)s]}{(C_0-\epsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s](1-\theta^n)}\left(\frac{k}{1-s}\right)^{n-1} \simeq \frac{(1-\theta^2)\theta^{n-1/2}}{(1-\theta^n)} \qquad (3.1)$$

$$\frac{(k+s)\theta}{(C_0-\epsilon-\frac{1}{\rho_1})[1-(1+s)\theta-(2k+1)s]} \simeq \frac{\theta^{2.5}}{(1-\theta)} \qquad (3.2)$$

Thus, to a good approximation, (2.6) becomes

Approximate Estimate I

$$|\mathrm{Re}[Z_0(\omega) - Z_0^{(n)}(\omega)]/\mathrm{Re}[Z_0(\omega)]| \leqq (\tilde{e}_n/\tilde{f}_n)(1+\tilde{e}_n/\tilde{f}_n) \qquad (3.3)$$

where

$$\tilde{e}_n = \left(\frac{(1-\theta^2)\theta^{n-1/2}}{1-\theta^n}\right)\left[1+\left\{\begin{array}{l}\left(\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right| - \frac{(1-\theta^2)\theta^{n-3/2}}{(1-\theta^{n-1})} - \frac{\theta^{2.5}}{1-\theta}\right)^{-1} \\ \times\left(\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right| - \frac{\theta^{2.5}}{1-\theta}\right)^{-1}\end{array}\right\}\right]$$

$$\tilde{f}_n = \left\{ \left| \operatorname{Im}\left[\frac{(-1)}{A(\omega-1)+\dfrac{(-1)}{A(\omega-2)-\dfrac{1}{\rho_1}}} + A(\omega) + \frac{(-1)}{A^*(-\omega-1)-\dfrac{1}{\rho_1}} \right] \right| \right.$$

$$- \left[\frac{(1-\theta^2)\theta^{n-3/2}}{1-\theta^{n-1}} + \frac{\theta^{2.5}}{1-\theta} \right] \left\{ 1 + \left| A(\omega-1)+\frac{(-1)}{A(\omega-2)-\dfrac{1}{\rho_1}} \right| \right.^{-1}$$

$$\left. \times \left[\left| A(\omega-1)+\frac{(-1)}{A(\omega-2)-\dfrac{1}{\rho_1}} \right| - \frac{(1-\theta^2)\theta^{n-3/2}}{1-\theta^{n-1}} - \frac{\theta^{2.5}}{1-\theta} \right]^{-1} \right\}$$

Next, our approximations transform (2.7) into

Approximate Estimate II

$$\left| \operatorname{Re}\left\{ Z_0(\tilde{\omega}) - \left[j\omega \left(\frac{(-1)}{A(\omega-1)+\dfrac{(-1)}{A(\omega-2)-\dfrac{1}{\rho_1}}} + A(\omega) + \frac{(-1)}{A^*(-\omega-1)-\dfrac{1}{\rho_1}} \right) \right]^{-1} \right\} \right| \qquad (3.4)$$

$$\lesssim (\tilde{e}/f)[1+\tilde{e}/f]$$

where

$$\tilde{e} = \left(\frac{\theta^{2.5}}{1-\theta} \right) \left[1 + \left| A(\omega-1) + \frac{(-1)}{A(\omega-2)-\dfrac{1}{\rho_1}} \right|^{-1} \left(\left| A(\omega-1) + \frac{(-1)}{A(\omega-2)-\dfrac{1}{\rho_1}} \right| - \frac{\theta^{2.5}}{1-\theta} \right)^{-1} \right]$$

and

$$f = \left| \operatorname{Im}\left(\frac{(-1)}{A(\omega-1) + \dfrac{(-1)}{A(\omega-2)-\dfrac{1}{\rho}}} + A(\omega) + \frac{(-1)}{A^*(-\omega-1)-\dfrac{1}{\rho_1}} \right) \right|$$

Further, (3.1) and (3.2) turn (2.9) into

Approximate Estimate III (Criterion for $\operatorname{Re}[Z_0(\omega)]$ to be negative)

$$\tilde{e}_n < \operatorname{Im}[L_n(\omega) + A(\omega) + L_n^*(-\omega)]$$

where

$$\tilde{e}_n = \left(\frac{(1-\theta^2)\theta^{n-1/2}}{1-\theta^n}\right)\left(1+\left\{\left(\left|A(\omega-1)+\frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right| - \frac{(1-\theta^2)\theta^{n-3/2}}{1-\theta^{n-1}} - \frac{\theta^{2.5}}{1-\theta}\right)^{-1} \times \left(\left|A(\omega-1) + \frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right| - \frac{\theta^{2.5}}{1-\theta}\right)^{-1}\right\}\right)$$

Then (2.10) becomes the formula below.

Approximate Estimate IV (Second criterion for $\mathrm{Re}[Z_0(\omega)]$ to be negative)

$$\tilde{e}'_n < \mathrm{Im}[L_n(\omega) + A(\omega) + L_n^*(-\omega)] \tag{3.5}$$

where

$$\tilde{e}'_n = \left(\frac{(1-\theta^2)\theta^{n-1/2}}{1-\theta^n}\right)\left\{1+|A(\omega-1)+L_{n-1}(\omega-1)|^{-1}\left(|A(\omega-1)+L_{n-1}(\omega-1)|-\frac{(1-\theta^2)\theta^{n-3/2}}{1-\theta^{n-1}}\right)^{-1}\right\}$$

Finally, (2.12) becomes

Approximate Estimate V (Sufficient condition for $\mathrm{Re}[Z_0(\omega)]$ to be negative)

$$\tilde{e} < \left|\mathrm{Im}\left[\frac{(-1)}{A(\omega-1) + \frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}} + A(\omega) + \frac{(-1)}{A^*(-\omega-1)-\frac{1}{\rho_1}}\right|\right. \tag{3.6}$$

where

$$\tilde{e} = \left(\frac{\theta^{2.5}}{1-\theta}\right)\left(1+\left|A(\omega-1) + \frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right|^{-1}\left(\left|A(\omega-1) + \frac{(-1)}{A(\omega-2)-\frac{1}{\rho_1}}\right| - \frac{\theta^{2.5}}{1-\theta}\right)^{-1}\right)$$

4. CONCLUSIONS

The theorems of section two are rigorous but quite complicated. If their use is required, they can be simply programmed for computer studies. The approximations of section three, however, are much more amenable to rough design calculations. Thus the theory of the preceding sections provides for a wide range of design needs.

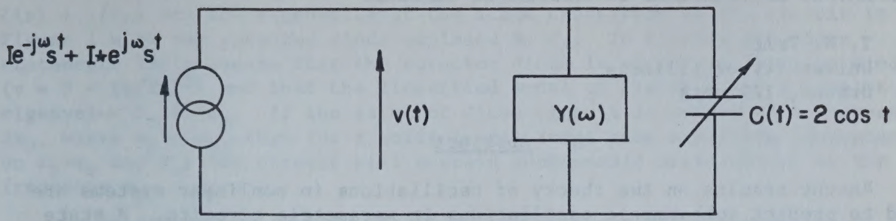

$Ie^{-j\omega_s t} + I{*}e^{j\omega_s t}$
v(t)
Y(ω)
C(t) = 2 cos t

Figure 1.

Normalized linear model for a parametric amplifier

REFERENCES

[1] B. J. Leon and J. V. Adams, Steady State Analysis of Linear Networks Containing a Single Sinusoidally Varying Element, Purdue University School of Electrical Engineering, TR EE66-4.

[2] O. Perron, Die Lehre von den Kettenbruchen, Teubner, 1957, vol. 2, Theorem 2.40.

SUBHARMONIC OSCILLATIONS IN PARAMETRIC CIRCUITS

T. N. Trick
University of Illinois
Urbana, Illinois

Abstract

Recent results on the theory of oscillations in nonlinear systems are used to predict subharmonic oscillations in parametric circuits. A state variable approach is used which allows one to extend nonlinear analysis beyond the usual second order nonlinear differential equation case. In particular, a fourth order network containing a varactor diode is considered. Conditions are arrived at such that this network will lock into subharmonic oscillations at two different frequencies. It is shown that the various pump frequencies at which the circuit can sustain subharmonic oscillations are related to the eigenvalues of the state transition matrix of a linearized model for the above nonlinear network. Experimental results are included which substantiate the theoretical work.

I. Introduction

The design of parametric amplifiers and frequency converters is carried out on the basis of a time-varying, finite frequency model. In actuality the parametric amplifier is a nonlinear system, not time-varying, and the number of frequencies present in the response is not just three or four, but an infinite number. Thus, quite often nonlinear phenomena, such as, frequency entrainment and subharmonic oscillations, occur in the real world circuit which are unexplainable on the basis of a time-varying model. Unfortunately, the nonlinear theorist has been unable to provide the parametric amplifier designer with better design criteria. The stability of unforced high order nonlinear systems has been discussed a great deal lately. However, a parametric amplifier must be stable to periodic inputs, and it has been shown [1] that zero input asymptotic stability does not imply stability for periodic inputs necessarily. Some stability results [2,3] and truncation error results [4] have been found for parametric amplifiers, but these results do not indicate how to design better parametric amplifiers.

It appears that a substantial contribution could be made to the design of parametric circuits if the cause of subharmonic oscillations could be pinpointed. At least we could then design so as to avoid subharmonic oscillations. In the next section a strong conjecture will be made concerning some necessary conditions for the existence of subharmonic oscillations. This conjecture will be supported by some recent theoretical results in the study of oscillations in nonlinear systems. Finally, an example will be presented along with experimental results.

II. Existence of Subharmonic Oscillations in Parametric Circuits

Consider the Thevenin equivalent of a varactor diode circuit illustrated in Figure 1. It is desired to know the frequencies at which this circuit will sustain subharmonic oscillations. The conjecture is as follows. The circuit in Figure 1 can be driven into subharmonic oscillation with a frequency of oscillation in the neighborhood of any of the finite complex zeros of the impedance function $Z(s) + 1/C_{\ell}s$, Figure 2, provided the zeros are sufficiently close to the $j\omega$ axis, i.e., the circuit loss is not too great. The constant C_{ℓ} is the capacitance of the varactor diode at its quiescent point. It should be noted that the zeros of $Z(s) + 1/C_{\ell}s$ are the eigenvalues of the state transition of the circuit in Figure 1 with the varactor diode replaced by C_{ℓ}. To clarify the above statements let's assume that the varactor diode is an abrupt junction diode $(v = \phi - (q/2k)^2)$ and that the linearized model of Figure 1 has a complex eigenvalue $\sigma_o + j\omega_o$. If the varactor diode circuit is pumped at a frequency $2\omega_p$, where $\omega_p \approx \omega_o$, then for a sufficiently large pump amplitude (dependent on $\omega_p - \omega_o$ and σ_o) the circuit will sustain subharmonic oscillations at the frequency ω_p.

In an attempt to justify the above conjecture a brief discussion of some recent results by Hale and others [5] in the theory of oscillations in nonlinear systems must be presented. Hale considers the differential equation

$$\dot{z}^* = Az^* + \epsilon Z(t,z^*,\epsilon) - \epsilon P_o Z(t,z^*,\epsilon) \ , \qquad (1)$$

$$A = \text{diag}\ (O_p, B) \ , \qquad (2)$$

$$Z = \text{col}(X,Y) \ , \qquad z = \text{col}(x,y) \ ,$$

where A is an nxn constant matrix; O_p is a pxp zero matrix; P_o is an averaging operator over the interval 0 to T; $0 \leq |\epsilon| \leq \epsilon_o$, $\epsilon_o > 0$; z, Z are n vectors and x, X are p vectors; t is periodic in t of period T, is continuous in t, z, ϵ, has a continuous first derivative with respect to ϵ and continuous first and second derivatives with respect to z for $-\infty < t < \infty$. The steady state solution of eq. (1) is given by the following iterates which converge for small enough ϵ.

$$z^{(o)} = \text{col}(a,0) = a^* \qquad (3)$$

$$z^{(k+1)}(t) = a^* + \epsilon \int^{\cdot} e^{A(t-u)}(I-P_o)Z(u,z^k(u),\epsilon)\,du \ . \qquad (4)$$

However, of actual interest is the equation

$$\dot{z} = Az + \epsilon Z(t,z,\epsilon) \ . \qquad (5)$$

It is shown that there exists a periodic solution of period T to eq. (5) if an $a(\epsilon)$ can be found such that

$$P_o Z(t, z, a(\epsilon), \epsilon) = 0 \ , \qquad (6)$$

or equivalently,

$$X_o(a_o,0,0) = \frac{1}{T}\int_0^T X(t,x,y,\epsilon)\,dt = 0 \ , \qquad (7a)$$

$$\det\left[\frac{\partial X_o(a_o,0,0)}{\partial x}\right] \neq 0 \ . \qquad (7b)$$

In order to determine the existence of steady state subharmonic oscillations the problem must first be formulated in the form of eq. (5). Assume that there exists a pair of complex conjugate eigenvalues $(-\sigma_1 + j\omega_1)$ and $(-\sigma_1 - j\omega_1)$ in the state transition matrix, and assume that σ_1 represents a small loss term which can be lumped with the nonlinear terms. Thus, the circuit equations can be put into the form

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdots \\ \dot{\underset{\sim}{y}} \end{bmatrix} = \left[\begin{array}{cc|c} j\omega_1 & 0 & 0 \\ 0 & -j\omega_1 & \\ \hline 0 & & B \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ \underset{\sim}{y} \end{bmatrix} + \epsilon \begin{bmatrix} X_1(t,x,y,\epsilon) \\ X_2(t,x,y,\epsilon) \\ \cdots\cdots \\ \underset{\sim}{Y}(t,x,y,\epsilon) \end{bmatrix}, \qquad (8)$$

where $\underset{\sim}{y}$ and $\underset{\sim}{Y}$ are n-2 vectors, and B is an n-2xn-2 constant matrix. Now suppose that the forcing function is periodic with period $2\pi/2\omega_o$, where $\omega_o \approx \omega_1$, and let

$$x_1 = z_1 e^{j\omega_o t}, \quad x_2 = z_2 e^{-j\omega_o t}.$$

Thus, eq. (8) can be written as

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \cdots \\ \dot{\underset{\sim}{y}} \end{bmatrix} = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & B \end{array}\right] \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ \underset{\sim}{y} \end{bmatrix} + \epsilon \left[\begin{array}{c} j\dfrac{(\omega_1-\omega_o)}{\epsilon} z_1 + e^{-j\omega_o t} X_1(z_1,z_2,\underset{\sim}{y},t,\epsilon) \\ -j\dfrac{(\omega_1-\omega_o)}{\epsilon} z_2 + e^{j\omega_o t} X_2(z_1,z_2,\underset{\sim}{y},t,\epsilon) \\ \hline \underset{\sim}{Y}(z_1,z_2,\underset{\sim}{y},t,\epsilon) \end{array}\right] \qquad (9)$$

Hence, the first iterate is

$$z_1^{(o)} = a, \quad z_2^{(o)} = \bar{a}, \text{ and } \underset{\sim}{y}^{(o)} = 0 .$$

Note that the last term in eq. (9) is periodic of period $2\pi/\omega_o$ satisfying Hale's conditions. The determining equations which indicate the possibility of the existence of a steady state subharmonic solution of period $2\pi/\omega_o$ are

$$P_o[j\frac{(\omega_1-\omega_o)}{\epsilon} a + e^{-j\omega_o t} X_1(a,\bar{a},0,t,\epsilon)] = 0 \qquad (10)$$

and

$$P_o[-j\frac{(\omega_1-\omega_o)}{\epsilon} \bar{a} + e^{j\omega_o t} X_2(a,\bar{a},0,t,\epsilon)] = 0 . \qquad (11)$$

If a solution to eqs. (10) and (11) exists and if eq. (7b) is satisfied, then we say that a subharmonic oscillation can exist in the circuit. We have not said anything about the stability of this solution, nor will we in this paper. Also, note that the nth order nonlinear differential equation reduces, by this iteration scheme, to a 2nd order problem.

III. An Example

In order to test our conjecture experimental and theoretical results were obtained for a fourth order nonlinear network. The linearized model of this circuit had two pair of complex conjugate zeros near the $j\omega$ axis, thus, the circuit should sustain subharmonic oscillations near either of the eigenvalues, depending on the pump frequency. The circuit under consideration is shown in Figure 3. The network equations for this circuit can be put in the form

$$\begin{bmatrix} \ddot{q}_{v_1} \\ \\ \ddot{q}_{v_2} \end{bmatrix} + \begin{bmatrix} \dfrac{q_{01}-q_{02}}{2k^2L_1} & \dfrac{-(q_{01}-q_{02})}{2k^2L_1} \\ \\ \dfrac{-(q_{01}-q_{02})}{2k^2L_2} & \dfrac{1}{L_2C_2}+\dfrac{q_{01}-q_{02}}{2k^2L_2} \end{bmatrix} \begin{bmatrix} q_{v_1} \\ \\ q_{v_2} \end{bmatrix} = \begin{bmatrix} \epsilon f_1+g \\ \\ \epsilon f_2 \end{bmatrix}, \quad (12)$$

where

$$g = \frac{E_p}{L_1}\cos 2\omega_o t \,, \qquad (13)$$

$$\epsilon f_1 = \frac{-(q_{v_1}-q_{v_2})^2}{4k^2L_1} - \frac{R_1}{L_1}\dot{q}_{v_1} \,, \qquad (14)$$

and

$$\epsilon f_2 = \frac{(q_{v_1}-q_{v_2})^2}{4k^2L_1} - \frac{R_2}{L_2}\dot{q}_{v_2} \,. \qquad (15)$$

Also,

$$q_1(t) = q_{v_1}(t) + q_{01} \,,$$

where q_{01} is the d-c component and k is determined by the particular varactor diode. Equation (12) must now be written in standard form, equation (5). First, a 2x2 matrix P must be found such that

$$\begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} + \begin{bmatrix} \omega_{r_1}^2 & 0 \\ 0 & \omega_{r_2}^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = P \begin{bmatrix} \epsilon f_1 + g \\ \epsilon f_2 \end{bmatrix} , \qquad (16)$$

where

$$P \begin{bmatrix} q_{v_1} \\ q_{v_2} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} .$$

Now let

$$x_1 = z_1 + u_1 \,,$$

where

$$\ddot{u}_1 + \omega_{r_1}^2 u_1 = P_{11} g \,,$$

and let

$$x_2 = z_2 + u_2 ,$$

where

$$\ddot{u}_2 + \omega_{r_1}^{2} u_2 = P_{21} g .$$

Thus, we have

$$\begin{bmatrix} \ddot{z}_1 \\ \ddot{z}_2 \end{bmatrix} + \begin{bmatrix} \omega_{r_1}^{2} & 0 \\ 0 & \omega_{r2}^{2} \end{bmatrix} = \epsilon \begin{bmatrix} P_{11} f_1 + P_{12} f_2 \\ P_{21} f_1 + P_{22} f_2 \end{bmatrix} = \epsilon \begin{bmatrix} F_1(t,u,z) \\ F_2(t,u,z) \end{bmatrix} . \quad (17)$$

Suppose now that one is interested in determining if a subharmonic oscillation exists near the frequency ω_{r_1}. Then, let $\omega_{r_1} \approx \omega_o$, and so,

$$\begin{bmatrix} \ddot{z}_1 \\ \ddot{z}_2 \end{bmatrix} + \begin{bmatrix} \omega_o^{2} & 0 \\ 0 & \omega_{r_2}^{2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \epsilon \begin{bmatrix} \dfrac{(\omega_o^{2} - \omega_{r_1}^{2})}{\epsilon} z_1 + F_1(t,u,z) \\ F_2(t,u,z) \end{bmatrix} . \quad (18)$$

Now as our first iterate we let

$$z_1^{(o)} = a e^{j\omega_o t} + \bar{a} e^{-j\omega_o t} \quad (19)$$

$$z_2^{(o)} = 0 . \quad (20)$$

An equivalent to the set of determining equations (10) and (11) is obtained by substituting eqs. (19) and (20) into eq. (18) and requiring the sum of the coefficients of the $e^{j\omega_o t}$ terms in the first row to be zero and requring the sum of the coefficients of the $e^{-j\omega_o t}$ terms in the first row to be zero. Substituting in the circuit values it was found that these determining equations did have a solution and thus subharmonic oscillations can exist near the frequency ω_{r_1}. A similar procedure can be carried out for investigating the existence of subharmonics near the frequency ω_{r_2}. One merely assumes that $\omega_o \approx \omega_{r_2}$ and proceeds from eq. (17) to eq. (18) with $\omega_{r_2}^{2}$ in the left side of eq. (18) replaced by ω_o^{2}, and the term on the right side of eq. (18) in the second row is

$$[\frac{\omega_o^{2} - \omega_{r_2}^{2}}{\epsilon} z_2 + F_2(t,u,z)] .$$

Again it is found that the subharmonics can exist at this frequency also.

The experimental results obtained showed that $f_{r_1} = 265$kHz and $f_{r_2} = 445$kHz. These values differed from our theoretical results considerably, $f_{r_1} = 160$kHz and $f_{r_2} = 900$kHz, primarily due to the fact that the stray capacitance of the inductors, which was appreciable, was neglected in the theoretical calculations. Since f_{r_1} = 250kHz the circuit was pumped in the neighborhood of 500kHz and it was found that for a sufficiently large pump amplitude the circuit did go into subharmonic oscillations of order 1/2 the pump frequency, Figure 4. The output voltage was measured across the varactor diode. Likewise, since f_{r_2} = 445kHz, the circuit was pumped near 890kHz. Again, for sufficiently large pump amplitudes, the circuit sustained subharmonic oscillations of order 1/2 the pump frequency, Figure 5.

IV. Concluding Remarks

Although the stability of the subharmonic oscillations were not investigated theoretically yet, the experimental results indicated that they were stable. In the future the region of stability should be investigated. Also, the accuracy of the first approximation should be investigated especially as the order of the nonlinear differential equation increases.

It is clear that no clear cut design procedure for parametric circuits has been found in this paper. However, it is hoped that by knowing what conditions in a high order nonlinear circuit cause subharmonic oscillations will enable one to design so as to at least avoid this anomaly.

V. Acknowledgments

The experimental results in this paper were obtained by Roger L. Fredrick. This work was supported by the Joint Services Electronics Program under Contract DA 28 043 AMC 00073(E), and by the Air Force Office of Scientific Research under Grant No. AFOSR 931-66.

References

1. C. A. Desoer, R. Liu, and L. V. Auth, Jr., "Linearity vs. Nonlinearity and Asymptotic Stability in the Large," IEEE Trans. on Circuit Theory, CT-12, p. 117, March 1965.

2. T. N. Trick, "Stability of Nonlinear Circuits with Periodic Inputs," Proceedings of the NEC, Vol. 22, pp. 912-916, 1966.

3. B. J. Leon and D. R. Anderson, "The Stability of Pumped Nonlinear Reactance Circuits," IEEE Trans. on Circuit Theory, Vol. CT-10, pp. 468-476, Dec. 1963 (Correction: Vol. CT-12, p. 143, March 1965).

4. D. R. Anderson and B. J. Leon, "Nonlinear Distortion and Truncation Errors in Frequency Converters and Parametric Amplifiers," IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 314-320, September 1965.

5. J. K. Hale, Oscillations in Nonlinear Systems, McGraw-Hill Book Co., 1963.

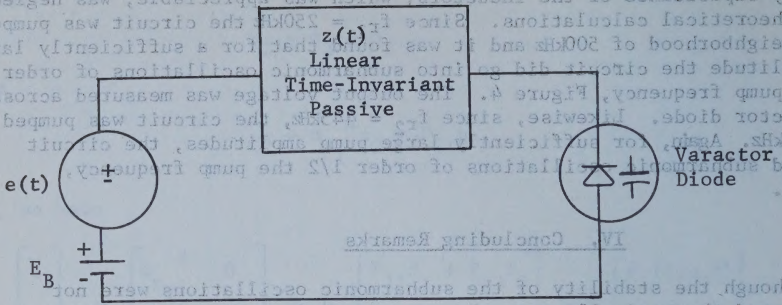

z(t)
Linear
Time-Invariant
Passive
e(t)
+
-
+
E_B
-
Varactor
Diode

Figure 1. Varactor Diode Circuit-Thevenin Equivalent

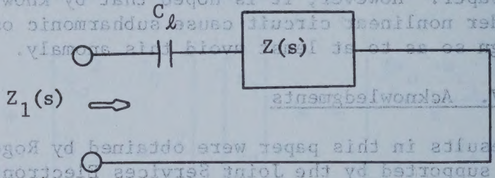


Figure 2. Complex Zeros of $Z_1(s)$ Determine Possible Subharmonic Oscillations

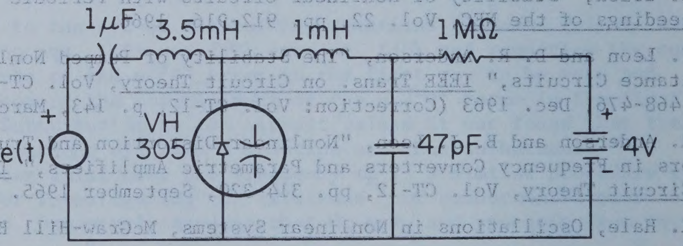

1μF
3.5mH
1mH
1MΩ
+
e(t)
VH
305
47pF
+
4V
-

Figure 3. Fourth Order Varactor Diode Circuit

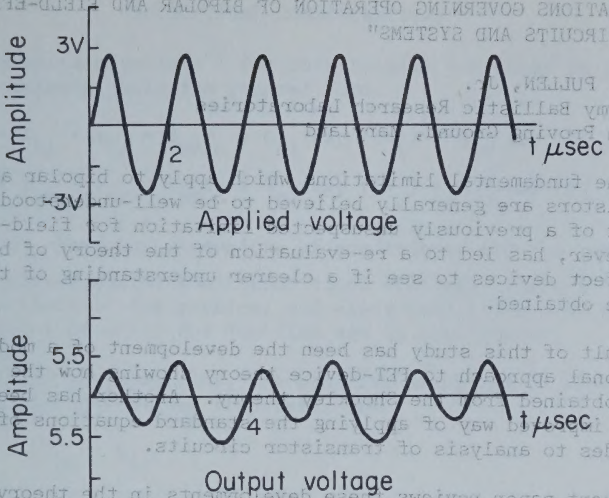


Figure 4. Subharmonic Oscillations when ω_o = 250kHz

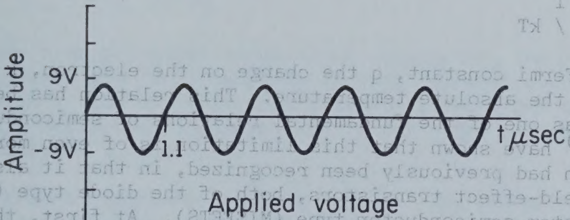


Figure 5. Subharmonic Oscillations when ω_o = 445kHz

"ON THE LIMITATIONS GOVERNING OPERATION OF BIPOLAR AND FIELD-EFFECT TRANSISTORS IN CIRCUITS AND SYSTEMS"

KEATS A. PULLEN, Jr.
U. S. Army Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland

Abstract: The fundamental limitations which apply to bipolar and field-effect transistors are generally believed to be well-understood. The recent establishment of a previously unsuspected limitation for field-effect devices1), however, has led to a re-evaluation of the theory of both bipolar and field-effect devices to see if a clearer understanding of their properties could be obtained.

One result of this study has been the development of a modification to the conventional approach to FET-device theory showing how the new limitation can be obtained from the Shockley theory. Another has been the development of an improved way of applying the standard equations of back-to-back junction diodes to analysis of transistor circuits.

The present paper reviews these developments in the theory of solid-state devices briefly and considers their implications in both the design of types of devices and the application of the devices in conventional and unusual circuit environments.

INTRODUCTION

The limitation applying to the transconductance gm per unit current I for both electron tubes and bipolar transistors can be expressed in terms of the equations:

$$\begin{aligned} gm &= \Lambda\ I \\ \Lambda &= q\ /\ kT \end{aligned} \qquad (1)$$

where Λ is the Fermi constant, q the charge on the electron, k the Boltzmann constant, and T the absolute temperature. This relation has been recognized for many years as one of the fundamental relations of semiconductor physics. Recent studies1) have shown that this limitation is of even more general application than had previously been recognized, in that it also applies to conventional field-effect transistors, both of the diode type (DIFETS) and the metal-insulator-semiconductor type (MISFETS). At first, this result appeared to be in disagreement with the theories of Shockley and Dacey-Ross, but re-examination of Shockley's considerations indicated that a transconductance limitation of the kind under discussion could be obtained from a change in Shockley's approximations.

This disclosure has interesting consequences in the basic theory of design of active circuits and networks, and it also has bearing on the design of both DIFET and MISFET devices. It further leads to new understanding on the behavior of semiconductor devices under exposure to the various kinds of radiation which might be encountered in outer space or in a nuclear environment.

In this paper we propose to review briefly the changes required in the Shockley theory to take account of the new limitation and to develop some of consequences which result therefrom. In particular, a broader concept of the behavior of all semiconductor type devices results which leads to a clearer understanding of and a more effective use of these devices.

THEORETICAL CONSIDERATIONS

The fundamental equations$^{2)}$ for back-to-back junctions in polarized semiconductor material take the general form:

$$\begin{aligned} I_b &= I_{bo} + I_{b1} + I_{b2} = G_{bo}V_o + G_{b1}\ \exp\ (\Lambda V_b) + G_{b2}\ \exp\ (\Lambda V_c) \\ I_c &= I_{co} + I_{c1} + I_{c2} = G_{co}V_o + G_{c1}\ \exp\ (\Lambda V_c) + G_{c2}\ \exp\ (\Lambda V_c) \end{aligned} \qquad (2)$$

where G_{bo} and G_{co} have the dimensions of admittance, and the remaining G-functions the dimensions of admittance times voltage. These G-functions are hyperbolic functions of the physical and electrical properties of the semiconductor material of which the junction set is constructed.

Now, in these two equations, the relative magnitude of the current components in a properly designed transistor are:

$$\begin{aligned} I_{b1} &\gg I_{bo} > I_{b2} \\ I_{c1} &\gg I_{co} > I_{c2} \end{aligned} \qquad (3)$$

As a result, if partial derivatives are taken of Eqs. 2, the results are:

$$\begin{aligned} y_i &= \partial I_b \ / \ \partial V_b = \Lambda I_{b1} \doteq \Lambda I_b \\ y_f &= \partial I_c \ / \ \partial V_b = \Lambda I_{c1} \doteq \Lambda I_c \\ y_r &= \partial I_b \ / \ \partial V_c = \Lambda I_{b2} \\ y_o &= \partial I_c \ / \ \partial V_c = \Lambda I_{c2} \end{aligned} \qquad (4)$$

Clearly, all the four principal conductance components, input, output, and forward and reverse, are functions of a specific current component and also the Fermi constant. The two most important ones, the input and the forward conductance, are determined approximately by the Fermi constant and the overall magnitudes of I_b and I_c.

Now, the transconductance per unit current observed with electron tubes at normal levels of current is significantly less than the 39,000 micromhos per milliampere specified by the Fermi constant, although with very well designed tubes it can be a significant fraction of this value. The actual emission and control is very similar in nature to that in the bipolar transistor, inasmuch as the electron emission is thermal emission against the work function of a solid-state boundary or barrier, and the grid control alters the boundary field. For this reason, it appeared that a similar kind of limitation should be present with all known kinds of field-effect devices. This was confirmed (in typical devices)$^{3)}$ by experimental measurements at very low values of channel current.

A study of Shockley's theory$^{4)}$ shows that in the Debye transition region at the boundary of the channel of the field-effect transistor, a potential jump of Λ^{-1}, or 26 millivolts, occurs. Shockley has assumed this change to be of negligible importance. Since, however, the physical length of the transition region is very small, a substantial voltage gradient can and does exist, and the increment may therefore be very important indeed.

In fact, it is possible, by proper selection of the doping distribution within the channel, to reduce to a very small value the voltage gradient or electric field required to bind the charges in the depletion region, that part of the channel in which charges are bound by the gate field and in which no channel current can flow. Under these conditions, the voltage jump across the transition region can certainly be important.

Measurements on typical devices show that this voltage jump is indeed important. When the device is biased toward cutoff, the condition of virtual channel closure is in fact the one in which transconductances unit per current approaching that specified by the Fermi constant do develop. Under these conditions with typical devices, the binding voltage can be substantial, several volts in fact, yet the limitation is clearly experienced, as is indicated by measurements on the device behavior.

Further consideration indicates that the closeness of approach of the actual behavior of the device to theoretical is dependent on the efficiency factor, K:

$$\partial I / I \partial y \simeq K \tag{5}$$

If the total current change equals the total current (as might occur at cutoff), then the value of K is unity. As the amount of total current increases compared to the change experienced, the value of K becomes smaller and smaller. This is in fact why relatively low values of K are obtained with electron tubes and presently available kinds of field-effect transistors at moderate values of device current. The change in current is too small compared to the total current flow at the specified bias, and the device is correspondingly inefficient.

The coefficient for the proportionality factor in Eq. 5 appears to depend on the derivative of a unit function at its point of discontinuity. Since this derivative is with respect to the boundary coordinate, and both the boundary and the unit function move in position together, this derivative of necessity is non-zero at the boundary. The same derivative of the unit function appears in the determination of the derivative of voltage with respect to the boundary coordinate, leading to a cancellation of the unit-function derivatives.

NOTES ON FET DESIGN

Examination of Eq. 5 indicates that the way to maximize the transconductance per unit current in a field-effect transistor is to increase the current change per unit current flow in the device. Physically, this means that the amount of current flow available as each new incremental channel is switched from the depleted state to the conducting state must be large compared to the current already flowing. For such a situation to exist, it is essential that the doping density be minimum at the center of the channel, and that it increase very rapidly in either direction from channel center. A hyperbolic cosine function of the general form:

$$\cosh \lambda y = (\exp \lambda y + \exp -\lambda y)/2 \tag{6}$$

is quite suitable as long as the value of λ is significantly larger than unity. Doping the semiconductor to lead to a maximizing of the value of λ leads to a device having a transconductance limit approaching the theoretical value. It is then easily shown that as y becomes vanishingly small, or it approaches Δy in magnitude, the relation:

$$\lim_{y \to 0} (\partial I / I \, \partial y) / (\partial V / \partial y) \rightarrow \Lambda \qquad (7)$$

Clearly, an hyperbolic or catenary distribution of dopant, with a large value of λ, can lead to a high practical transconductance per unit current.

APPLICATION NOTES

Now that it is clear that the transfer admittances for both bipolar and field-effect transistors are dependent on the Fermi constant, and the input admittance of the bipolar transistor is also dependent on this constant, it is of interest to examine the consequences of these relations on circuit design.

The Fermi constant, as defined in Eq. 1, is dependent on three parameters, the charge on the electron, q, Boltzmann's constant, k, and the absolute temperature, T, in degrees Kelvin. Two of these parameters are natural "constants", and the third is a variable in the mathematical sense. This latter variable is the only one over which the designer has any control, and available semiconductor devices operate in accord with theory over a limited range of this temperature only. The range of variation of the Fermi "constant" with variations of temperature from -20° C to +100° C is from an increase of 19 percent to a decrease of 21 percent with respect to the 39,000 value. Clearly, this is well within the tolerances on such parameters of a typical bipolar transistor as its beta.

As is clear from Eqs. 4, both the input conductance and the forward conductance can be specified to tolerances set by base and collector current measurements and by the environmental temperature. Conventional transistor stabilization techniques stabilize the static collector current, and thereby stabilize the forward transfer properties of the device at the same time. The input, or base current, for the bipolar transistor, is then a function of the diffusion length of the minority carriers in the base region for the transistor. As a consequence, it is not possible to stabilize the input current or the input admittance for the bipolar transistor, and it becomes necessary to provide for a range of values for these variables.

If the signal source impedance for the transistor amplifier is sufficiently small, the calculated characteristics of the amplifier need to be corrected only for the effect of base-spreading resistance. Otherwise design may be based on the intrinsic transconductance. The applicable equations are:

$$y_f = y_{f'} / (1 + y_{i'} r_{b'}) = \Lambda \, I_c / (1 + \Lambda \, I_b \, r_{b'})$$

$$y_i = y_{i'} / (1 + y_{i'} r_{b'}) = \Lambda \, I_b / (1 + \Lambda \, I_b \, r_{b'}) \qquad (8)$$

$$K_v = y_f \, Z_L$$

where the equation for K_v, the voltage gain, assumes that the source and the load impedancesare sufficiently small that the denominator terms involving them may be neglected.

The correction for base-spreading resistance may be rather small, or it may be in some cases rather large, but it is normally stable, and once the correction has been made, further adjustments are not required.

The current flowing to the base of the bipolar transistor is an inverse-square function of the minority carrier lifetime in the base region. This lifetime is at best rather unstable, and it varies as a function of voltages, currents, temperature, and previous life history. Both the input admittance and this division ratio, I_c/I_b, vary together, as they must for Eq. 4 to hold.

CONCLUSIONS

The result of use of Eq. 4 is that a simple and direct approach to the design of transistor circuits is obtained which is based on the inherent nonlinear properties of the devices. It leads directly to usable designs behaving as predicted by theory. Stable parameters are used for primary design, the less stable ones being allowed for by proper marginal design techniques. And the techniques apply with both bipolar and field-effect transistors, and also to circuits in which they are operated in high-radiation environments as well as in conventional environments.

REFERENCES

1) BRL Memorandum Report 1746, "A New Limitation in Field-Effect Transistor Operation," U. S. Army Ballistic Research Laboratories, 1966.

2) These equations are simplified versions of those given in standard texts such as Middlebrook.

3) Evans, L., and Pullen, K. A., Jr., "Limitation of Properties of Field-Effect Transistors," Proc. IEEE, January 1966 (Letter to the Editor).

4) Shockley, W., "A Unipolar "Field-Effect" Transistor," Proc. IRE, Nov. 1952.

DESIGN THEORY OF A NEGATIVE RESISTANCE AMPLIFIER USING A TRANSISTORIZED ISOLATOR

R.S.GOPALAN and B.A.SHENOI
Department of Electrical Engineering
University of Minnesota,Minneapolis. Minn.

ABSTRACT

An isolator using transistors has been realized from the gyrator circuit developed by Shenoi and an analysis of the two-port network parameters valid for a larger frequency range than that assumed for the gyrator circuit is presented. It is shown that the cumbersome admittance parameters of the device could be simplified under some constraints on the network elements and a simple model containing only a few parasitic elements around one ideal gyrator can be obtained. Analysis of a nonreciprocal negative resistance amplifier containing this nonideal isolator and a tunnel diode and exhibiting a Butterworth response is presented. The gain-bandwidth product for the amplifier is shown to be restricted not only by the RC product of the tunnel diode but also the values of the parasitic elements of the isolator. A procedure for synthesizing the amplifier whose gain-bandwidth product approaches that of an amplifier containing an ideal isolator and experimental evidence using the transistorized isolator circuit to support the theory has also been briefly outlined.

I INTRODUCTION

The theory of nonreciprocal negative resistance amplifiers embedded with a circulator and a tunnel diode has been extensively studied by several authors (1,2,3), since the invention of the tunnel diode. Analysis of a nonreciprocal amplifier employing a two-port isolator has also been briefly treated(4). In all cases,however, the authors have assumed an ideal circulator or isolator working over such a large range of frequency that their theoretical results are more optimistic than that obtainable in practice. In this paper an attempt has been made to realize a nonideal isolator built by using presently available transistors and modify the existing design theory of nonreciprocal amplifiers to include the effect of the frequency response characteristics of the nonideal isolator.

The basic circuit used for realizing the isolator is the transistor circuit of a gyrator developed by Shenoi(5). In its analysis, he assumed the low frequency T equvivalent circuit for the transistors which is valid over a very limited frequency range compared to their cut-off frequency. Hence thoough his circuit approximated an ideal gyrator over the frequency range he considered its limitations beyond this frequency range were not studied.

II REALIZATION OF AN ISOLATOR

In our present work, the hybrid π equivalent circuit of the transistors is chosen for deriving the s.c.admittance parameters of Shenoi's gyrator circuit which is shown in Fig.1(without its associated biassing and coupling circuitry) so as to extend the frequency range over which the device behavior is better described by these y parameters or a 2-port equivalent circuit model. Even after making some simplifying assumptions in derivig the y parameters,it is found that the 2 transfer admittances y_{12} and y_{21} of the gyrator circuit given by (1) and (2) below are different from each other.

$$y_{12} = \frac{-g_{12}}{1+\frac{px_2}{\omega_{T2}}} \tag{1}$$

$$y_{21} = \frac{g_{21}}{1+\frac{px_1}{\omega_{T2}}} \tag{2}$$

where

$$\frac{1}{g_{21}} = \frac{R_{BB1}}{\beta_1} + r_{e1} + R_{E1} \tag{3}$$

$$\frac{1}{g_{12}} = \frac{1}{\alpha_{03}}\left[R_4 + \left(1+\frac{R_4}{R_{E2}}\right)\left(r_{e2}+\frac{R_{BB2}}{\beta_2}\right) + \left(1+\frac{R_4}{R_5}\right)\left(r_{e3}+R_{BB3}\,\overline{1-\alpha_{03}}\right)\right.$$
$$\left. + \left(r_{e2}+\frac{R_{BB2}}{\beta_2}\right)\left(r_{e3}+R_{BB3}\,\overline{1-\alpha_{03}}\right)\left(\frac{R_{E2}+R_4+R_5}{R_{E2}R_5}\right)\right] \tag{4}$$

$$x_1 = \frac{(G_{BB1}+G_{E1})\,g_{e1}}{G_{BB1}(g_{e1}+G_{E1}) + \frac{G_{E1}g_{e1}}{\beta_1}} = g_{21}(R_{BB1}+R_{E1}) \tag{5}$$

$$x_2 = \frac{g_{12}}{\alpha_{03}}\left[R_{BB2} + \frac{R_{BB3}+R_4}{u} + \frac{R_4R_{BB2}}{R_{E2}} + \frac{R_4R_{BB3}}{R_5u}\right.$$
$$\left. + R_{BB2}\left(r_{e3}+R_{B3}\,\overline{1-\alpha_{03}}\right)\left(\frac{R_{E2}+R_4+R_5}{R_{E2}R_5}\right)\right] \tag{6}$$

$$u = \frac{\omega_{\alpha3}}{\omega_{T2}} = \frac{R_{E2}+R_4+R_{BB3}\left(\frac{R_{E2}+R_4+R_5}{R_{E2}R_5}\right)}{R_{E2}+R_4+r_{e3}\left(\frac{R_{E2}+R_4+R_5}{R_{E2}R_5}\right)} \tag{7}$$

p = complex frequency variable

β_i = common emitter sc current gain of transistor T_i (i=1,2,3) at low frequency

α_{oi} = low frequency common base sc current gain of T_i

$\omega_{\alpha i}$ = alpha cutoff frequency of T_i

ω_{Ti} = radian frequency at which common emitter current gain of T_i is unity

$R_{BBi} = r_{bbi} + R_{ext}$ where r_{bbi} is the base resistance of T_i and R_{ext} is the external resistance connected to the base.

R_{Ei} = External resistance connected to the emitter terminal of T_i

Hence it is necessary to satisfy the following two conditions (8) and (9) which lead to the requirements given by (10) and (11). But from (7) we note that $u > 1$ and also $\omega_{\alpha 3} \gg \omega_{T2}$. From (3)-(6), we find that the condition given by (8) can be satisfied by adjusting the values of R_{E1} and R_4 whereas (9) can be satisfied by adjusting the values of R_{BB1} and R_{BB2}, provided that R_{E1} and R_4 are small compared to R_{BB1} and R_{BB2}. This interesting property of the gyrator is used to satisfy (12), in which case the gyrator becomes antireciprocal.

$$g_{12} = g_{21} \tag{8}$$

$$\frac{\omega_{T1}}{x_1} = \frac{\omega_{T2}}{x_2} = \omega_1 \tag{9}$$

$$R_{E1} + r_{e1} + \frac{R_{BB1}}{\beta_1} = \frac{1}{\alpha_{03}}\Bigg[R_4 + \left(1 + \frac{R_4}{R_{E2}}\right)\left(r_{e2} + \frac{R_{BB2}}{\beta_2}\right) + \\ + \left(1 + \frac{R_4}{R_5}\right)\left(r_{e3} + R_{BB3}\,\overline{1-\alpha_{03}}\right) + \left(r_{e2} + \frac{R_{BB2}}{\beta_2}\right)\left(r_{e3} + R_{BB3}\,\overline{1-\alpha_{03}}\right) \\ \left(\frac{R_{E2} + R_4 + R_5}{R_{E2}\,R_5}\right)\Bigg] \tag{10}$$

$$\frac{\omega_{T1}}{R_{E1} + R_{B1}} = \frac{\omega_{T2}\,\alpha_{03}}{\left[R_{BB2} + \frac{R_{BB3}}{u} + \frac{R_4}{u} + \frac{R_4\,R_{BB2}}{R_{E2}} + \frac{R_4}{R_5}\frac{R_{BB3}}{u} + \\ + R_{BB2}\left(r_{e3} + R_{BB3}\,\overline{1-\alpha_{03}}\right)\frac{R_{E2} + R_4 + R_5}{R_{E2}\,R_5}\right]} \tag{11}$$

$$y_{12} = -y_{21} = \frac{-g_{12}}{1 + \left(\frac{p}{\omega_1}\right)} \tag{12}$$

Next we connect a two-terminal RL network of admittance equal to y_{21} across the input and output terminals of the gyrator. The overall y matrix of the unilateral circuit reduces to (13). As the third step in obtaining a simplified isolator circuit, the output s.c, admittance y_{22} shown in (13) is converte to a pure conductance G_{22} by connecting a complementary admittance given by (14) across the output terminals. The s.c. input admittance y_{11} in (13) has a zero on the negative real axis given by (15). So we make $G_{L3}x_1 \ll g_{12}$ in order that this zero is very close to $-\omega_{T1}$. Now if the passband of the amplifier is confined to about $(\omega_{T1}/5)$, the zero can be considered as nondominant and hence neglected. Under this one restriction, the final y matrix of the isolator reduced to a simple model by properly choosing the values of R_{BB1} and R_{BB2} etc. is given by (16) and the equivalent circuit by Fig. 2 . In this simplified model, the pole of the admittance parameters is within the passband of the amplifier and therefore the design theory would be so developed as to include effect of the frequency-dependant behavior of the two-port isolator.

$$\begin{bmatrix} G_{L3} + \dfrac{g_{12}\left(\dfrac{1}{\beta_1} + \dfrac{1}{\omega_{T1}} + 1\right)}{1 + \dfrac{p}{\omega_1}} & \dfrac{-2g_{12}}{1 + \dfrac{p}{\omega_1}} \\ 0 & G_{L1} + \dfrac{g_{12}\left(\dfrac{1}{\beta_2} + \dfrac{c}{\omega_{T2}} + 1\right)}{1 + \dfrac{p}{\omega_1}} \end{bmatrix} \qquad (13)$$

where

$$c = 1 + \frac{R_4}{R_{E2}} + \left(r_{e3} + R_{BB3}\,\overline{1-\alpha_{03}}\right)\frac{R_{E2} + R_4 + R_5}{R_{E2}R_5}$$

$$\frac{\left(\dfrac{G_{22} - G_{L1}}{\omega_1} - \dfrac{g_{12}c}{\omega_{T2}}\right)p}{1 + \dfrac{p}{\omega_1}} \qquad (14)$$

whereby $\quad G_{22} = G_{L1} + g_{12}\left(1 + \frac{1}{\beta_2}\right)$

$$-\left(\frac{G_{L3} + g_{12} + \dfrac{g_{12}}{\beta_1}}{G_{L3}x_1 + g_{12}}\right)\omega_{T1} \qquad (15)$$

$\approx -\omega_{T1}$ when $G_{L3}x_1 \ll g_{12}$, since $x_1 > 1$; $\beta_1 \gg 1$

$$\begin{bmatrix} \dfrac{g_{12}}{1+\frac{p}{\omega_1}} & \dfrac{-2g_{12}}{1+\frac{p}{\omega_1}} \\ 0 & G_{22} \end{bmatrix} \qquad (16)$$

Since there is transmission only from port 2 to port 1,the generator is to be connected at port 2 and the load at port 1. The nonreciprocal amplifier network in which are connected the isolator,the generator with an internal resistance R_g, the load R_L, the tunnel diode represented by the parallel connection of a negative resistance $-R_D$ and a capacitance C_D, and the lossless equalizer network N to be designed for getting the desired frequency response is shown in Fig.3.

III DESIGN THEORY OF THE AMPLIFIER

The transducer gain G_T of the amplifier circuit shown in Fig.3 is derived by using the scattering matrix analysis,for the case $R_L > R_D$, in which the normalization at port 1 and 2 is done with respect to $1/g_{12}$ and $\hat{R}_L = \left|\frac{R_L R_D}{R_L - R_D}\right|$ respectively. The transducer gain is given by (17) in which we substitute for $|S_{22}|^2$ from (18) in order to get a Butterworth response for the transducer gain,as

$$G_T = \frac{4 R_g \frac{1}{g_{12}} \left(\frac{R_D}{R_L - R_D}\right)}{\left(R_g + \frac{1}{g_{22}}\right)^2 g_{22}^2} \; \frac{|S_{21}|^2}{|S_{22}|^2} \qquad (17)$$

$$|S_{22}|^2 = S_{22}(p)\, S_{22}(-p)\Big|_{p=j\omega} = \frac{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}{K_n + \left(\frac{\omega}{\omega_c}\right)^{2n}} \qquad (18)$$

indicated by (19). After factorizing (18), $S_{22}(p)$ is formed such that it has poles and zeros only in the left half p-plane and the corresponding output driving point admittance at port 2 is evaluated. This admittance,however, has to be realized in such a way that the 'first' element at port 2 is a shunt capacitance of value no less than C_D and the 'last' element is a series combination of inductance of value no less than $(1/\omega_1 g_{12})$ and a resistance exactly equal to $1/g_{12}$ so that the parasitic elements of the isolator as well as the shunt capacitance of the tunnel diode can be absorbed by the elements of the equalizer network.

$$G_T = \frac{4 R_g \frac{1}{g_{12}}}{\left(R_g + \frac{1}{g_{22}}\right)^2 g_{22}^2} \; \frac{(K_n - 1)}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}} \qquad (19)$$

The constraints give rise to the following inequalities

$$k_n \leq \left(1 + \frac{2 \sin \pi/2n}{\hat{R}_L C_d \omega_c} \right)^{2n} \tag{20}$$

$$\frac{1}{g_{12}} = \hat{R}_L \frac{k_n^{\frac{1}{2}} - 1}{k_n^{\frac{1}{2}} + 1} \tag{21}$$

$$\frac{\left(k_n^{\frac{1}{2}} - 1\right)}{\left(k_n^{\frac{n-1}{2n}} - 1\right)} \omega_c \leq \frac{\omega_1}{\sin \pi/2n} \tag{22}$$

The inequality (20) due to the tunnel diode parasitic element has already been obtained by Youla and Smilen(1). The inequality (22) is due to the isolator. Hence any tunnel diode amplifier using this isolator has to satisfy both (20) and (22) for gain-bandwidth requirement and, in addition, n has to be even. The parasitic inductance due to the isolator can however be absorbed in the equaliser without affecting the bandwidth. In (22), as n is increased, the right hand side tends to zero and the left hand side tends to ω_c and hence by using a higher order filter, the restriction on bandwidth can be overcome, provided ω_c is within $\omega_{t_1}/5$ or $\omega_{t_2}/5$, whichever is smaller. The d.c. gain is fixed completely by $\hat{R}_L$ and $\frac{1}{g_{12}}$ and for a given $\hat{R}_L$, $\frac{1}{g_{12}}$ can be chosen suitably to give any gain.

IV EXPERIMENTAL RESULTS

A non-reciprocal 4th order amplifier was designed to give a power gain of 13 db and a bandwidth of 300 kc/s. The values of the various elements are $R_L = 600$, $R_g = 75$, $-R_d = -100$ and $C_a = 100$ pf. The isolator to be used in the circuit was required to have the following y-matrix

$$\begin{bmatrix} \dfrac{0.01}{1 + \dfrac{p}{1.25 \times 10^6}} & \dfrac{-0.02}{1 + \dfrac{p}{1.25 \times 10^6}} \\ 0 & 0.0133 \end{bmatrix}$$

The transistors used were of the alloy junction type having the following characteristics

$f_{t_1} = 2.77$ Mc/s $\quad f_{t_2} = 2.28$ Mc/s $\quad f_{a_3} = 4.2$ Mc/s

$r_{e_1} = 25\ \Omega$ $\quad r_{e_2} = 25\ \Omega$ $\quad r_{e_3} = 25\ \Omega$

$r_{bb_1} = 120\ \Omega$ $\quad r_{bb_2} = 115\ \Omega$ $\quad r_{bb_3} = 128\ \Omega$

at $V_{CE} = -6$ V and $I_c = -1$ mA. The isolator that was built was found to have the following y-matrix

$$\begin{bmatrix} \dfrac{0.0117}{1+\dfrac{p}{1.17\times10^{6}}} & -\dfrac{0.0189}{1+\dfrac{p}{1.17\times10^{6}}} \\ 0 & 0.0129 \end{bmatrix}$$

The theoretical and the experimental response curves are shown in Fig. 4. The circuit of the isolator used in the experiment is shown in Fig. 5.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the discussion with Professor S. C. Dutta Roy. They also wish to thank Mr. John Viner for his help in the experiments and Mr. Yogish Alekal for his help in preparing the diagrams.

The support of the above research work was received through an NSF Research Assistantship granted to the first author and an NSF Grant (no. GP - 3017) granted to the second author.

REFERENCES

[1] D. C. Youla and L. I. Smilen, "Optimum negative resistance amplifiers", Proc. Symp. on Active Networks and Feedback Systems, PIB, MRI, pp. 214-319, 1960.

[2] E. W. Sard, "Gain bandwidth performance of maximally flat negative conductance amplifiers", ibid., pp. 319-345, 1960.

[3] E. S. Kuh and J. D. Paterson, "Design Theory of optimum negative resistance amplifiers", Proc. IRE, vol. 49, pp. 1043-1051, June 1961.

[4] H. J. Carlin and D. C. Youla, "Network synthesis with negative resistance", Proc. Symp. on Active Networks and Feedback Systems, PIB, MRI, pp. 27-69, 1960.

[5] B. A. Shenoi, "Practical realization of a gyrator circuit and RC gyrator filters", IEEE Trans. on Circuit Theory, vol. CT-12, pp. 374-381, September 1965.

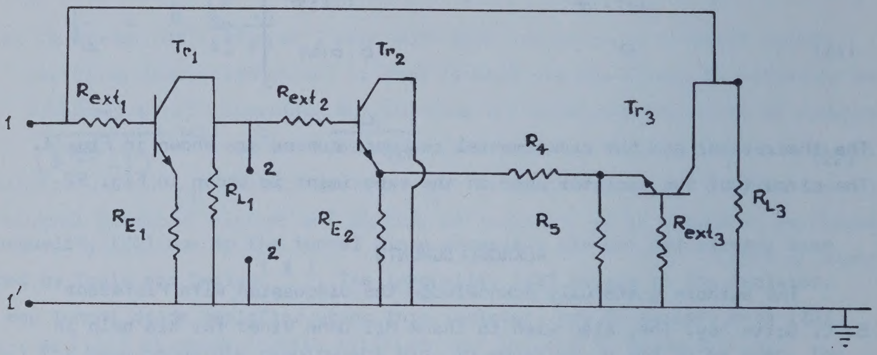

Tr1
Tr2
Tr3
Rext1
Rext2
R4
R5
Rext3
RL1
RL3
RE1
RE2
1
1'
2
2'

FIG. 1

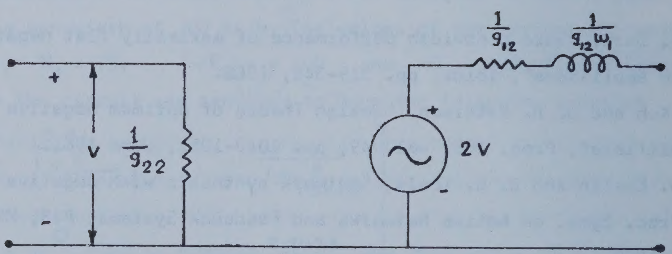

+
-
V
1/g22
1/g12
1/g12ω1
+
-
2V

FIG. 2

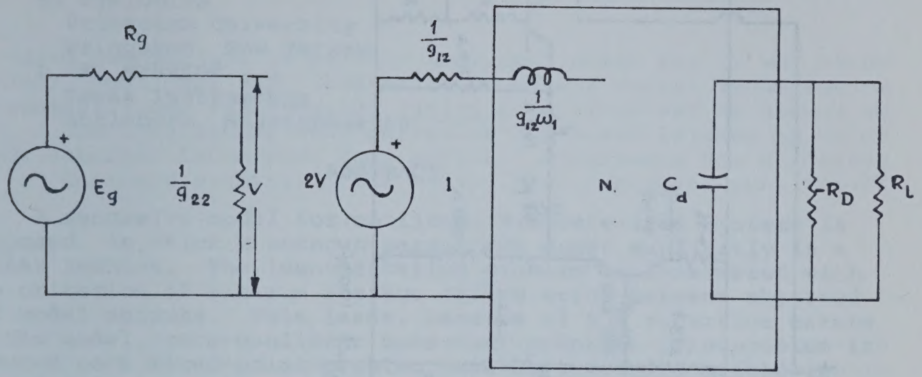
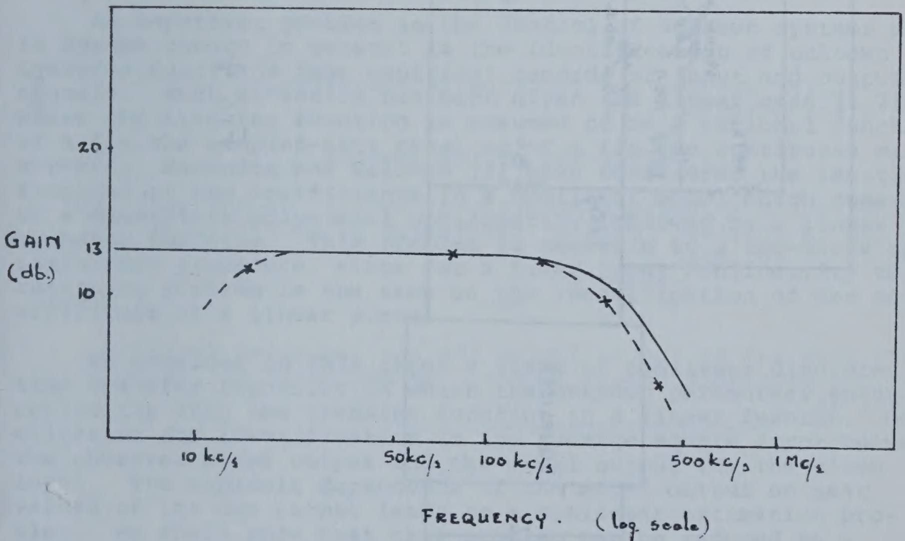

Rg
+
Eg
1/g22
V
1/g12
1/g12ω1
+
2V
I
N.
Cd
RD
RL

FIG. 3



FREQUENCY RESPONSE OF AMPLIFIER

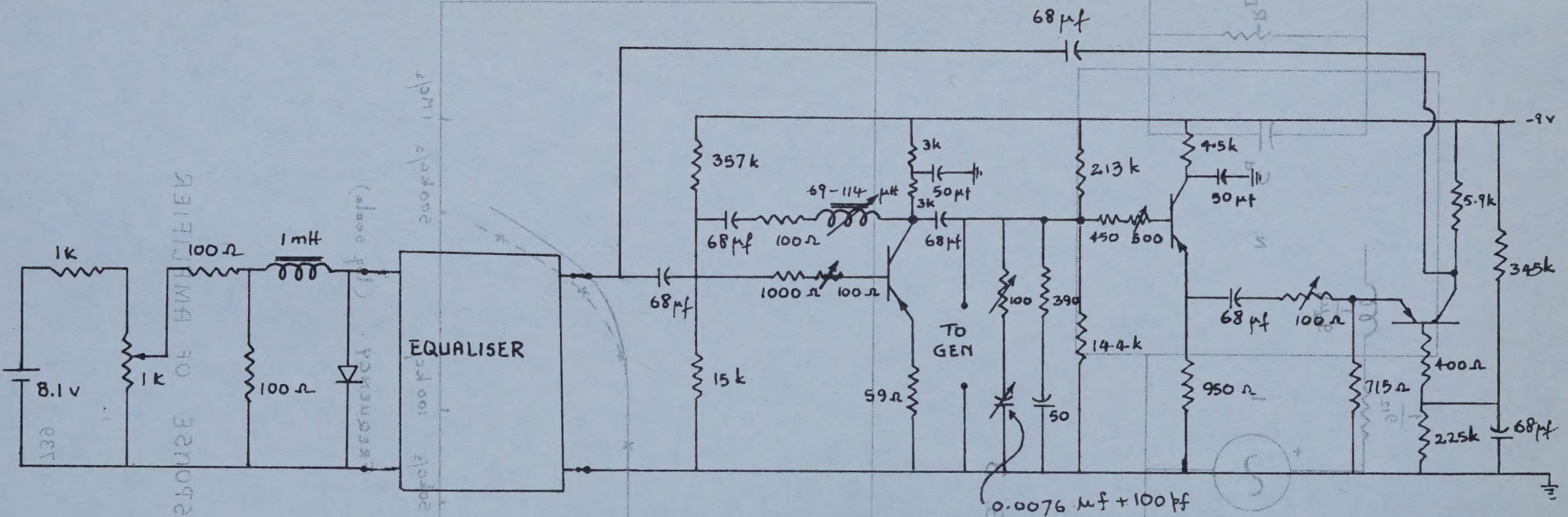

68 μf
-9V
1K
8.1 V
1K
100 Ω
1mH
100 Ω
EQUALISER
68 μf
357k
68 μf
100 Ω
69-114 μH
1000 Ω
100 Ω
15k
3k
50 μf
3k
68 μf
59 Ω
TO GEN
100
390
50
0.0076 μf + 100 pf
213 k
450
500
144k
4.5k
50 μf
950 Ω
68 μf
100 Ω
715 Ω
5.9k
345k
400 Ω
225k
68 μf

Fig 4

IDENTIFICATION OF NONLINEAR SAMPLED-DATA SYSTEMS

K. STEIGLITZ
Princeton University
Princeton, New Jersey
L. E. McBRIDE
Texas Instruments
Attleboro, Massachusetts

ABSTRACT

A recursive model for nonlinear discrete-time systems is proposed, in which K unknown parameters enter explicitly in a linear fashion. The identification problem is considered with the criterion of minimum average square error between observed and model outputs. This leads, because of the recursive nature of the model, to a nonlinear numerical problem. The problem is reduced to a fixed-point problem, and an iterative method is proposed for solution which depends for its convergence on the selection of a gain parameter. Under additional assumptions it is proved that the iteration procedure is locally stable for a sufficiently small gain parameter. Computational examples are presented which illustrate the practicality of the method.

I. INTRODUCTION

An important problem in the control of unknown systems and in system theory in general is the identification of unknown transfer functions from empirical records of input and output signals. Much attention has been given the linear case [1,2], where the transfer function is assumed to be a rational function of z (in the sampled-data case) or of s (in the continuous case). Recently, Narendra and Gallman [3] have considered the identification of the coefficients in a nonlinear model which consists of a memoryless polynomial nonlinearity followed by a linear transfer function. This problem is amenable to a two-stage optimization procedure, since for a fixed input nonlinearity the resulting problem is the same as the identification of the coefficients of a linear plant.

We consider in this paper a class of nonlinear discrete-time transfer functions in which the unknown parameters enter explicitly into the transfer function in a linear fashion. Our criterion for identification is the average square error between the observed plant output and the model output for the given input. The explicit dependence of the model output on past values of its own output leads to a nonlinear estimation problem. We shall show that this problem can be reduced to a fixed-point problem which can be solved in a practical way by iterative methods.

II. THE MODEL

Let x_n and y_n be the input and output at time n of a discrete-time system. We assume that observed records of x_n and a noise-corrupted y_n called w_n are available over some time interval. Call the model output v_n and assume for the model the recursive transfer function

$$v_n = \sum_{j=1}^{K} c_j f_j(x_n, x_{n-1}, \ldots, x_{n-M+1}, v_{n-1}, \ldots, v_{n-L}) \tag{1}$$

where the f_j are known functions and the c_j are unknown parameters whose values are to be estimated. The functions f_j must be chosen on the basis of a priori information; there appears to be no general basis for selecting them except physical reasoning and experience. Notice that this model includes the linear time-invariant case, where the difference equation

$$v_n = \sum_{j=1}^{M} c_j x_{n-j+1} + \sum_{j=M+1}^{M+L} c_j v_{n-j+M} \tag{2}$$

determines a rational function of z. It is evident, on the other hand, that wide varieties of dynamic behavior can be obtained by choosing the f_j to be nonlinear functions of the past and present inputs, and past outputs.

Since the x_n will be fixed and the same for both the model and the plant, we will abbreviate (1) by writing

$$v_n = \sum_{j=1}^{K} c_j f_{jn}(v) \quad , \tag{3}$$

where $f_{jn}(v)$ denotes the function f_j evaluated at time n for past plant outputs v.

III. STATEMENT OF THE PROBLEM

With the v_n defined recursively by (3), we can state our problem as follows: Find $\underline{c}$, the vector of coefficients $c_1, \ldots, c_K$, so that

$$\overline{[w_n - v_n]^2} = \text{min.} \tag{4}$$

where here and in the following the bar indicates averaging over the available record length. We may write (4) in more detail as

$$\overline{\left[w_n - \sum_{j=1}^{K} c_j f_{jn}(v)\right]^2} = \text{min.} \tag{5}$$

This is the average square error between observed and predicted plant outputs where the model has access to its own past output only. Since v depends on $\underline{c}$, the derivatives of (5) with respect to the c_j are not linear functions of the c_j. Hence the minimization problem cannot be solved directly by matrix inversion.

IV. THE LINEAR ESTIMATE

A related problem which does lead to equations that are linear in the c_j is the following:

$$\overline{\left[w_n - \sum_{j=1}^{K} c_j f_{jn}(w)\right]^2} = \min. \tag{6}$$

Here the model uses observed past outputs instead of its own past outputs. Since the output will in general be corrupted by noise, the two criteria (5) and (6) will lead to different estimates even if the model is appropriate. That the estimate based on (5) can result in a considerably more accurate identification than that based on (6) has been shown by example in the linear transfer function case [4], and will be shown by example in the nonlinear case later on. Equation (6) can be solved easily for $\underline{c}$, however, yielding

$$\underline{c} = \underline{\underline{D}}^{-1}\underline{b} \quad , \tag{7}$$

where $\underline{\underline{D}}$ is a KxK matrix with elements

$$\overline{f_{in}(w) f_{jn}(w)} \quad , \qquad i,j = 1,\ldots,K \quad , \tag{8}$$

and b is a K-vector with components

$$b_i = \overline{f_{in}(w) w_n} \quad , \qquad i = 1,\ldots,K \quad . \tag{9}$$

This estimate for $\underline{c}$ will be called the *linear estimate*.

V. REDUCTION TO A FIXED-POINT PROBLEM

We now consider the problem of solving (4) exactly. Differentiating (4) with respect to c_i we obtain

$$\overline{(w_n - v_n)\,\frac{\partial v_n}{\partial c_i}} = 0 \quad , \qquad i = 1,\ldots,K \quad , \tag{10}$$

as necessary conditions for an optimal choice of the c_i. For any given $\underline{c}$ we may calculate successively v_n, $f_{jn}(v)$, and $\partial v_n/\partial c_i$. We shall write these as explicit functions of $\underline{c}$ as follows:

$$\left.\begin{aligned} & v_n(\underline{c}) \quad , \\ & f_{jn}(\underline{c}) = f_{jn}(v(\underline{c})) \quad , \\ \text{and} \quad & \gamma_{in}(\underline{c}) = \partial v_n(\underline{c})/\partial c_i \quad . \end{aligned}\right\} \tag{11}$$

Equation (10) becomes

$$\overline{(w_n - v_n(\underline{c}))\,\gamma_{in}(\underline{c})} = 0 \quad , \qquad i = 1,\ldots,K \quad , \tag{12}$$

or

$$\overline{\left[w_n - \sum_{j=1}^{K} c_j f_{jn}(\underline{c})\right]\gamma_{in}(\underline{c})} = 0 \quad , \quad i = 1,\ldots,K \, . \qquad (13)$$

Solving for the c_j that occur linearly, we have

$$\underline{c} = \underline{\underline{\Delta}}^{-1}(\underline{c})\underline{\beta}(\underline{c}) \quad , \qquad (14)$$

where $\underline{\underline{\Delta}}(\underline{c})$ is the KxK matrix with elements

$$\Delta_{ij} = \overline{\gamma_{in}(\underline{c}) f_{jn}(\underline{c})} \quad , \qquad (15)$$

and $\underline{\beta}$ is the K-vector with components

$$\beta_i = \overline{\gamma_{in}(\underline{c}) w_n} \quad . \qquad (16)$$

Writing

$$\underline{\psi}(\underline{c}) = \underline{\underline{\Delta}}^{-1}(\underline{c})\underline{\beta}(\underline{c}) \qquad (17)$$

(14) becomes

$$\underline{\psi}(\underline{c}) = \underline{c} \quad . \qquad (18)$$

Thus the solution to our original identification problem can be written as the solution to the fixed-point problem for the vector function $\underline{\psi}(\underline{c})$.

VI. CALCULATION OF THE GRADIENT SIGNAL γ_{in}

The gradient signal γ_{in} can be calculated on a digital computer by taking small finite differences in c_i and subtracting the resulting signals v_n. This has been found to be overly sensitive to roundoff errors and inadequate for iterative solutions of (18). A more satisfactory method is to obtain the signal γ_{in} by filtering. To see how to do this write

$$v_n = \sum_{j=1}^{K} c_j f_{jn} \quad . \qquad (19)$$

Differentiating with respect to c_i, this becomes

$$\frac{\partial v_n}{\partial c_i} = f_{in} + \sum_{\ell=1}^{L} g_{\ell n} \frac{\partial v_{n-\ell}}{\partial c_i} \quad , \qquad (20)$$

where

$$g_{\ell n} = \sum_{j=1}^{K} c_j \frac{\partial f_{jn}}{\partial v_{n-\ell}} \quad . \qquad (21)$$

Thus γ_{in} is obtained as the output of the time-varying linear filter defined by (20) with input f_{in}. This method of gradient calculation has proved quite adequate for an iterative solution.

VII. ITERATIVE SOLUTION OF THE FIXED-POINT PROBLEM

A natural method for the solution of the fixed-point problem (18) is the substitution iteration; i.e. we choose a starting value for $\underline{c}$, say $\underline{c}^o$, and find by iterative calculation

$$\underline{c}^{\nu} = \underline{\psi}(\underline{c}^{\nu-1}) \quad , \quad \nu = 1,2,3,\ldots \quad . \tag{22}$$

This is equivalent to substituting for the exact past output of the model a past output computed on the basis of an assumed coefficient vector, and finding linear estimates for a new coefficient vector. If this converges to some coefficient vector $\underline{c}^{\infty}$, (18) and hence (13) is satisfied and the original estimation criterion will be met. The iteration (22) can be viewed as a discrete-time system with state $\underline{c}$ and transition function $\underline{\psi}$. A sufficient condition for the local stability of (22) is that $\underline{\psi}$ be a contraction mapping in some sphere about its fixed point. This is equivalent to the norm in K-space being a Lyapunov function for the discrete system. An analytical answer to the question of stability is not feasible, since $\underline{\psi}$ is an extremely complicated function. The practical answer to the question is "no". This can be seen from the results of the following simple computational experiment:

A random signal x_n was filtered by a known transfer function to obtain an output y_n. No noise was added to y_n, so that $w_n = y_n$. The initial value $\underline{c}^o$ was chosen to be the known correct value for the transfer function involved, so that the minimum average square error was zero. The resulting sequence $\underline{c}^{\nu}$ hovered about $\underline{c}^o$ for a few iterations, but finally diverged violently. Evidently computational noise disturbed the system and caused it to leave an unstable equilibrium point.

It is apparent that in a small region about its fixed point, the system (22) can be viewed as a linear control system, and that it may be unstable. This suggests adding a parameter G which plays the role of the gain of the linearized system. With this end, the following iteration rule has been used:

$$\underline{c}^{\nu} = G\,\underline{\psi}(\underline{c}^{\nu-1}) + (1-G)\underline{c}^{\nu-1} \quad , \quad 0 < G < 1 \quad . \tag{23}$$

This moves the coefficient vector in the same direction as (22), but moves it only a fraction G of the distance. Notice that (23) is equivalent to using a direct substitution algorithm for the function

$$\underline{\varphi}(\underline{c}) = G\,\underline{\psi}(\underline{c}) + (1-G)\underline{c} \quad , \quad 0 < G < 1 \quad , \tag{24}$$

which has the same fixed-points as $\underline{\psi}$.

The question arises as to whether we can always choose a gain G sufficiently small so as to ensure that a fixed-point of $\underline{\psi}$ is a stable fixed-point of $\underline{\varphi}$. It is shown elsewhere [6] that this is true under reasonable conditions for $\underline{\psi}$, provided that we generalize $\underline{\varphi}$ slightly by defining the i-th component of $\underline{\varphi}$ as

$$\varphi_i = G\,\delta_i \psi_i + (1-G\delta_i)c_i \quad , \tag{25}$$

where $\delta_i = \pm 1$ and depends on whether $\partial\psi_i/\partial c_i$ is less than or greater than 1 in the region around the fixed point. In practice, it was never found necessary to use negative δ_i's and the $\underline{c}$ of (24) was always found to be stable for sufficiently small G. Usually satisfactory convergence was obtained for G in the range .01 to .15. It should be noted that choosing too small a G will result in highly damped behavior and slow convergence. Because of this, it is desirable to start with a relatively large G and reduce it until stable convergence is obtained.

In the next section we present the results of some automatic computations which illustrate various aspects of the iterative method. The linear estimate (7) was used as a starting point for these iterations; this has proved to be a satisfactory method to obtain a reasonable $\underline{c}^o$. In each example one-hundred points of a white, gaussian, zero-mean, unit-variance signal x_n were filtered through a fixed plant to produce an output signal y_n. White gaussian noise was added to y_n to produce a corrupted output signal w_n, which was used in the identification procedure. The normalized average square noise is defined by

$$\overline{N^2} = \overline{(w_n-y_n)^2} \;/\; \overline{y_n^2} \;, \tag{26}$$

and the normalized average square error is defined by

$$\overline{E_1^2} = \overline{(w_n-v_n)^2} \;/\; \overline{y_n^2} \;, \tag{27}$$

and is the quantity to be minimized.

Computation time for these examples was about .5 to 1.5 seconds per iteration on the IBM 7094, depending on the model order.

VIII. COMPUTATIONAL EXAMPLES

Example 1 - The Effect of Gain on Stability for a Linear Transfer Function Identification Problem

Actual plant: $y_n = .5x_n + .5x_{n-1} + 1.2726y_{n-1} - .81y_{n-2}$

Model: $v_n = c_1x_n + c_2v_{n-1} + c_3v_{n-2}$

Gain: Run 1 $G = 0.1$ Run 2 $G = 0.15$

N_2: 0.4

Final $\underline{c}$: Run 1 $c_1 = 0.7238$, $c_2 = 1.347$, $c_3 = -0.8684$

Final $\overline{E_1^2}$: Run 1 .4700

The results of the two runs, with the same input and noise samples, and with gains .1 and .15 are shown in Figure 1. It is evident that the procedure is stable for the smaller gain and unstable for the larger gain. Such a dependence of the behavior on the gain is typical. The final converged value of the normalized average square error is .4700, which is more than the average square noise because the model cannot exactly represent the actual transfer function.

Example 2 - Product Feedback with an Exact Model

Actual plant: $y_n = .5x_n + .9y_{n-1} + x_n y_{n-1}$

Model: $v_n = c_1 x_n + c_2 v_{n-1} + c_3 x_n v_{n-1}$

Gain: $G = 0.03$

$\overline{N^2}$: Run 1 0.4, Run 2 0.2

Final $\underline{c}$: Run 1 $c_1 = 0.4546$, $c_2 = 0.9058$, $c_3 = 0.9317$

Run 2 $c_1 = 0.4661$, $c_2 = 0.9039$, $c_3 = 0.9504$

Final $\overline{E_1^2}$: Run 1 0.3744, Run 2 0.1869

This example illustrates the behavior of the identification method for two different noise levels, an exact model, and a plant with a feedback term which is the product of the present input with the previous output. As seen in Figure 2, the convergence is not monotonic except at the final stages, showing that the region of stability of the iterative method extends outside of the sphere within which it is contractive.

Example 3 - Identification of a Unity-Feedback System with Saturation Nonlinearity

Actual plant: $y_n = \mathrm{sat}(x_{n-1} - y_{n-1})$, where $\mathrm{sat}(z) = \begin{cases} -1 & z < -1 \\ z & |z| < 1 \\ 1 & z > 1 \end{cases}$

Model: Run 1 $v_n = \sum_{j=1}^{2} c_j (x_{n-1} - v_{n-1})^j$

Run 2 $v_n = \sum_{j=1}^{3} c_j (x_{n-1} - v_{n-1})^j$

Gain: .1

$\overline{N^2}$: .5

Final $\underline{c}$: Run 1 $c_1 = 0.6728$, $c_2 = -0.09655$

Run 2 $c_1 = 0.9185$, $c_2 = -0.4366$, $c_3 = -0.07214$

Final $\overline{E_1^2}$: Run 1 0.7120, Run 2 0.4781

This example illustrates the identification of a nonlinear system with an approximate model. The effect of the models is to approximate the saturation nonlinearity with a polynomial of order K. Since the saturation nonlinearity is an odd function, the cubic term is important, and the third order model provides a much better fit than the second order model. Figure 3 shows the normalized average square error, $\overline{E_1^2}$, for both runs. Also shown is

$$\overline{E_2^2} = \overline{(y_n - v_n)^2} \,/\, \overline{y_n^2} ,$$

which is a measure of the accuracy of the identification. Figure 4 shows plots of the functions

$$f(x) = \sum_{j=1}^{K} c_j x^j \quad , \quad K = 2,3 \quad , \tag{28}$$

using the final values of coefficients obtained from the two runs, illustrating how the model fits polynomial approximations to the saturation nonlinearity.

IX. CONCLUSIONS

The recursive nonlinear model (1) is not, of course, a completely general nonlinear model, nor is it "best" in any sense. It does have, however, some distinct advantages. First, it includes the linear time-invariant case where the transfer function is a rational function of z, or is a pulse transfer function derived from a rational function of s. In fact, with this model the iterative method presented here provides a new method for the identification of linear transfer functions, as an alternative to those described in references [1,2,4]. Second, it is sufficiently general to exhibit, with the proper choice of the f_j, wide varieties of nonlinear behavior. Third, the appearance of the parameters of the model in an explicit linear fashion makes possible a fairly general digital computer program for transfer function identification, and the functions f_j can be specified and changed quite independently of the iteration algorithm.

The examples described in the preceding section are typical of the experiments performed by the authors in testing the iterative method. In no case was there any difficulty in achieving convergence with a positive gain parameter and all the $\delta_i = +1$, although local stability of the iteration was proved only under assumptions about the mapping $\underline{\psi}$ and with the introduction of the sign parameters δ_i. Future investigations might consider in more detail the mapping $\underline{\psi}$ and the derivatives $\partial\psi_i/\partial c_i$ at a fixed-point. Also of interest are methods for automatic gain selection and methods for the acceleration of convergence, such as Aitken's δ^2 process [5].

X. ACKNOWLEDGMENT

This work was supported by U.S. Army Research Office - Durham under Contract DA-31-124-ARO-D-292, and by the National Science Foundation Grant NSF GK-270. The computer facilities used were supported in part by Grant NSF GP-579.

REFERENCES

1. R. E. Kalman, "Design of a self-optimizing control system", Trans. ASME, Vol. 80, pp. 468-478, February 1958.

2. M. J. Levin, "Estimation of a system pulse transfer function in the presence of noise", IEEE Trans. on Automatic Control, Vol. AC-9, pp. 229-235, July 1964.

3. K. S. Narendra and P. G. Gallman, "An iterative method for the identification of nonlinear systems using a Hammerstein model", to be published.

4. K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems", IEEE Trans. on Automatic Control, Vol. AC-10, pp. 461-464, October 1965.

5. P. Henrici, "Elements of Numerical Analysis", (Book), John Wiley & Sons, New York, 1964. Chapters 4 and 5.

6. K. Steiglitz and L. E. McBride, "Identification of Nonlinear Sampled-Data Systems", Technical Report No. 14, Communications Laboratory, Department of Electrical Engineering, Princeton University, Princeton, N.J., June 1966. See Appendix.

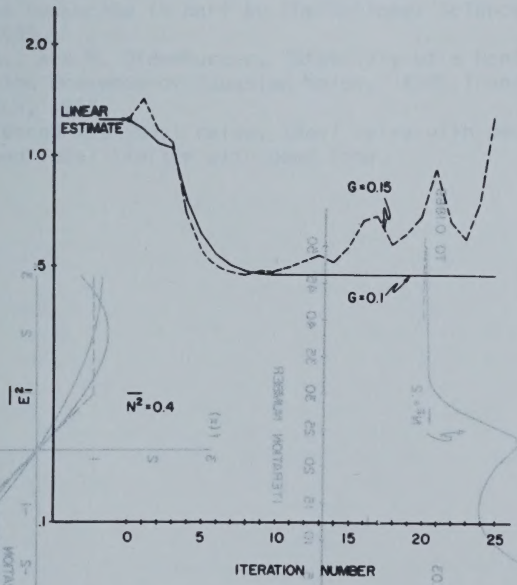


Fig. 1

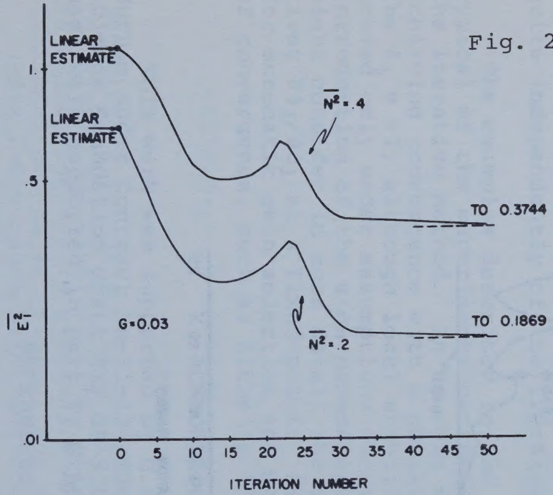


Fig. 2

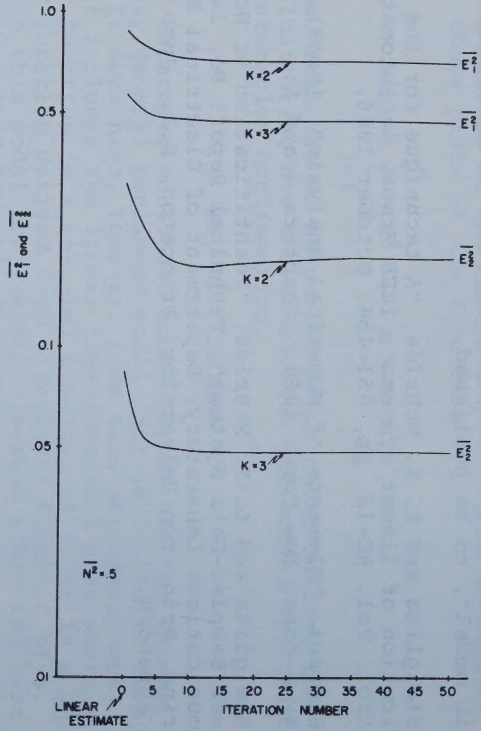


Fig. 3

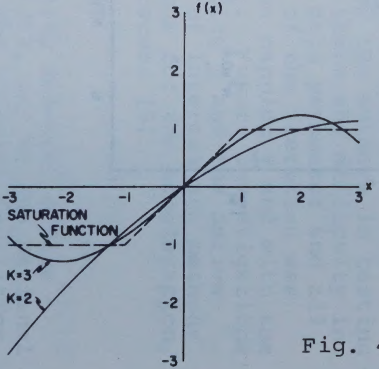


Fig. 4

SELF-OSCILLATION OF NONLINEAR SYSTEMS WITH GAUSSIAN NOISE*

S. H. WU and J. A. HEINEN
Department of Electrical Engineering
Marquette University
Milwaukee, Wisconsin

Recently, a stability criterion for nonlinear systems with a class of one single instantaneous nonmemory saturating type nonlinearity subject to Gaussian noise has been provided by Sridhar and Oldenburger[1] using the statistical linearization technique and the equivalent gain concept. The nonlinearities considered in their paper belong to a very restricted class#

In this paper, a general equivalent gain formula in the presence of Gaussian noise for an arbitrary symmetric saturating type nonlinearity is derived. This type of nonlinearity is important since it can be used to approximate any nonmemory instantaneous saturating nonlinear element to the desired degree. Thus it can be used to study the stability and oscillation in the presence of Gaussian noise of any feedback systems containing a saturating instantaneous nonmemory nonlinearity for which it might otherwise be impossible to obtain an analytical equivalent gain expression. Also, as a degenerated case, an expression of equivalent gain for an n^{th} order quantizer is derived. The expressions derived are general in scope and can be easily calculated with the aid of a digital computer.

As an example, the self-oscillation of a pulse-width modulated system in the presence of Gaussian noise is studied.

*This work was supported in part by The National Science Foundation under Grant No. GK-855

1. Sridhar, R., and R. Oldenburger, "Stability of a Nonlinear Feedback System in the Presence of Gaussian Noise," ASME Trans. Ser. D, Vol. 84, no. 1, March, 1962.

# They considered the ideal relay, ideal relay with dead zone, ideal limiter, and ideal limiter with dead zone.

ANALYSIS OF DELTA-MODULATED FEEDBACK SYSTEMS*

S. H. WU and A. N. MICHEL
Department of Electrical Engineering
Marquette University
Milwaukee, Wisconsin

Because of the simplicity and economy of delta modulation, this form of modulation has been given a great deal of attention in communication systems.[1,2,3,4] Very recently, a delta modulated feedback system has been studied from a control engineer's point of view.[5,6] Murphy and Shuraym[6] have shown that the delta modulated feedback system can be made to be stable in the sense of Lagrange and it is impossible to be made stable asymptotically in the Liapunov sense. Thus the delta modulated system will always have sustained oscillations if not unstable.

In this paper, we shall first show that delta modulated systems can be considered as a special degenerated type of pulse-width modulated system then by using this equivalence, a discrete model of delta modulated system will be given. From this model, all modes of oscillation can be found by an extension of the method given by Jury and Nishimura.[7] An example will be given to illustrate the technique developed.

---

* This work was supported in part by The National Science Foundation under Grant No. GK-855

1. Shouten, J. F., de Jager and J. A. Greefkes, "Delta Modulation, a New Modulation System for Telecommunication", Phillips Technical Review, Vol. 13, pp. 237-245, March 1952.
2. Libois, L. J., "Un nouveau Procede de modulation codee, la modulation en Δ", Onde Electrique, Jan. 1952, p. 26.
3. Nakanaru, Yoshimasa and H. Kaneko, "Delta Modulation Encoder", NEC Research and Development, Oct. 1960, pp. 46-54.
4. Inose, Hiroshi, Yasuhiko, and J. Murakal, "A Communication System by Code Modulation-Δ-Σ Modulation", J. IECE, Japan, Vol. 44, No. 11, pp. 1775-1780, November 1961.
5. Halijak, C. A. and J. S. Tripp, "A Deterministic Study of Delta Modulation", Proc. JACC, 1963, pp. 376-385.
6. Murphy, G. J. and G. Shuraym, "Delta Modulation in Feedback Control Systems" Proc. NEC, 1965, pp. 609-612.
7. Jury, E. I. and T. Nishimura, "On the Periodic Modes of Oscillation in Pulse-Width Modulated Systems," Trans. ASME Ser. D. J. of Basic Eng. March, 1962, pp. 71-84.

THE CONJUGATE GRADIENT MINIMIZATION METHOD IN FUNCTION SPACE

A.D. WAREN,* L.S. LASDON,** S.K. MITTER***

ABSTRACT

The method of conjugate gradients, an efficient means of solving static optimization problems, is extended to problems of optimal control. Convergence is shown to be more rapid than with the method of steepest descent, yet only function and gradient evaluations are required. The directions of search generated by the method are always directions of descent, so its stability properties are superior to those of second variational techniques.

The performance of the method is linked to its properties when applied to quadratic problems, which are reviewed here. Three numerical examples are given.

1. INTRODUCTION

A general formulation of the optimal control problem involves both terminal constraints on the state variables and inequality constraints on the state and control variables enforced along the entire trajectory. Penalty functions have often been used to convert such problems to a sequence of "unconstrained" problems, i.e., ones with no terminal or inequality constraints. [see ref. 1,2,3,4] It is evident that the efficiency of these methods depends greatly on the technique used to solve the unconstrained optimal control problem. Presently available techniques all have shortcomings. The convergence of steepest descent methods is often slow [see ref. 1] whereas second-variational and Newton methods may not converge at all. Thus there is strong motivation for developing more efficient techniques for solving unconstrained optimal control problems.

Similar difficulties existed, until recently, in the field of finite dimensional optimization, i.e., mathematical programming. However, in the past few years several rapidly convergent finite dimensional unconstrained minimization techniques have been developed. Among these are the method of Fletcher and Powell [see ref. 5] and the Fletcher-Reeves [see ref. 6] adaptation of the conjugate gradient method of Hestenes and Stiefel [see ref. 7]. Both these procedures generate conjugate directions of search and therefore minimize a positive definite quadratic function of n variables in n steps. In addition the directions generated are always directions of descent and thus, for relatively smooth functions, the function value is decreased at each step.

---

* Fenn College of Engineering, Cleveland State University, Cleveland, Ohio.

** Division of Organizational Sciences, Operations Research Group and Systems Research Center, Case Institute of Technology, Cleveland, Ohio.

*** Systems Research Center, Case Institute of Technology, Cleveland, Ohio.

The combination of these properties implies that the methods converge rapidly to the nearest local minimum for a general function of n variables. Experience has shown that both techniques converge much more rapidly, in general, than the method of steepest descent while requiring only function and gradient evaluations. Moreover, their stability properties are superior to those of second order Newton methods, and second order derivatives are not required.

Function space analogues of the steepest descent and second order Newton Techniques have been developed and applied to problems of optimal control. In particular Kelly and Bryson, [see ref. 8,9] Mitter, [see ref. 10] and others have developed steepest descent and second order methods. However, the analogues of the conjugate direction techniques have not yet been forthcoming. Since these methods are generally acknowledged to be the most powerful presently available for finite dimensional minimization problems, it seems appropriate to consider their generalization to optimal control.

This paper describes an extension of the Fletcher-Reeves [see ref. 6] conjugate gradient method to function space problems. The computational simplicity of this algorithm led to its selection. As in the steepest descent method the gradient trajectory must be computed and stored. In addition the conjugate gradient technique requires the computation of the norm of the gradient and the storage of one other trajectory, the actual direction of search.

Despite its simplicity, computational results illustrate its marked superiority to the method of steepest descent. These results are substantiated by theoretical developments.

It is shown here that the directions in function space generated by the conjugate gradient method are such that the objective function is decreased at each step. This leads to the results of section 3 concerning convergence from arbitrary starting points as the number of iterations tends to infinity. The rate of convergence is best investigated by considering the performance of the algorithm on a problem with a linear system and quadratic objective (Section 4). Here the results of Antosiewicz and Rheinboldt [see ref. 11] apply, showing that the directions of search are conjugate and hence the function is minimized over an expanding sequence of sets. In addition the distance from the current estimate of the solution to the optimal point decreases monotonically.

Further results in the linear-quadratic case have been obtained. For linear-quadratic problems in one state variable the conjugate gradient method achieves a value of the objective at least as low as that obtained by the method of steepest descent at each iteration (assuming the same starting point for both). Also, a class of quadratic problems is exhibited for which the conjugate gradient method finds the optimal solution in a finite number of steps.

None of the above properties are shared by the method of steepest descent, and account for the rapid convergence of the conjugate gradient method for quadratic problems. This implies that for those general nonlinear problems which may be approximated by linear-quadratic problems near the optimum, the convergence is rapid.

2. CONJUGATE GRADIENT ALGORITHM

2.1 Problem Formulation

Consider the following problem:

$$\text{minimize} \qquad J = \phi(\underline{x}(t_f)) \tag{2.1}$$

$$\text{subject to} \quad \dot{\underline{x}} = \underline{f}(\underline{x},\underline{u},t) \tag{2.2}$$

$$\underline{x}(t_o) = \underline{c} \tag{2.3}$$

where $\underline{x}$ is an n vector, $\underline{u}$ is an m vector, and t_0, t_f are fixed. It is assumed that given a control, u, (2.2) and (2.3) can be solved for a unique $\underline{x} = \underline{x}(\underline{u})$, and thus $J = J(\underline{u})$ is a function of $\underline{u}$ alone. Further more the existence of the gradient of $J(u)$, $\nabla J(\underline{u}) \equiv g(\underline{u})$ is assumed. As mentioned previously, the objective function (2.1) may include penalty function terms to account for constraints.

For the balance of the paper only the case of a single control function $u(t)$ $(m=1)$ will be considered. The extension to the multicontrol case is straightforward.

2.2 Algorithm

The conjugate gradient algorithm requires the computation of the gradient trajectory. Let

$$H = \sum_{i=1}^{n} \lambda_i f_i \tag{2.4}$$

where

$$\dot{\lambda}_i = - \sum_{j=1}^{n} \lambda_j \frac{\partial f_j}{\partial x_i} \tag{2.5}$$

$$\lambda_i(t_f) = \frac{\partial \phi}{\partial x_i}\bigg|_{t=t_f} \qquad i = 1, \ldots, n \tag{2.6}$$

Then the gradient is

$$g(u) = \frac{\partial H}{\partial u} \tag{2.7}$$

Let $u_i(t)$ be the i^{th} approximation to the optimal control $u^o(t)$. The corresponding gradient $g(u_i)$ is computed by solving the state equations (2.2) - (2.3) forwards with $u = u_i$, solving the adjoint equations (2.5) - (2.6) backwards and then computing $g(u_i)$ from (2.7).

The algorithm proceeds as follows:

$$u_0 = \text{arbitrary} \tag{2.8}$$

$$g_0 = g(u_0) \tag{2.9}$$

$$s_0 = - g_0 \tag{2.10}$$

$$\text{Choose} \quad \alpha = \alpha_i \quad \text{to minimize} \quad J(u_i + \alpha s_i) \tag{2.11}$$

$$u_{i+1} = u_i + \alpha_i s_i \tag{2.12}$$

$$g_{i+1} = g(u_{i+1}) \tag{2.13}$$

$$\beta_i = (g_{i+1}, g_{i+1})/(g_i, g_i) \tag{2.14}$$

$$s_{i+1} = - g_{i+1} + \beta_i s_i \tag{2.15}$$

where

$$(g_i, g_j) = \int_{t_o}^{t_f} g_i(t) g_j(t) dt \tag{2.16}$$

Note that the new direction of search, s_{i+1} is not the negative gradient direction, $- g_{i+1}$ but is computed via (2.15). The distance travelled in this direction is determined by the one dimensional minimization in (2.11).

3. CONVERGENCE

Let the control u be an element of a Hilbert space H and $J(u)$ a Frechet differentiable mapping from H to the real numbers. The conjugate gradient method when applied to $J(u)$ generates directions s_i which are always directions of descent, i.e.,

$$\frac{d}{d\alpha} J(u_i+\alpha s_i)_{\alpha=0} < 0 \tag{3.1}$$

and this guarantees that $J(u)$ is decreased at each step. These statements are proved below.

Theorem 1

If $g(u_i) = g_i \neq 0$ then $(S_i, g_{i+1}) = 0 \ \forall\ i$ and $\frac{d}{d\alpha} J(u_i+\alpha s_i)|_{\alpha=0} = (s_i, g_i) = -\,||g_i||^2 \forall i$.

Proof

Let $\alpha = \alpha_i$ minimize $J(u_i+\alpha s_i)$. Then

$$\frac{d}{d\alpha} J(u_i+\alpha s_i)|_{\alpha=\alpha_i} = (s_i, g_{i+1}) = 0 \tag{3.2}$$

From (2.15)

$$(g_i, s_i) = (g_i, -g_i+\beta_{i-1}s_{i-1}) \tag{3.3}$$

$$= -\,||g_i||^2 + \beta_{i-1}\,(g_i, s_{i-1}) = -\,||g_i||^2 \tag{3.4}$$

Theorem 2

If $g_i \neq 0$ then $J(u_{i+1}) < J(u_i)$

Proof

Assume there exists no $\alpha > 0$ such that

$$J(u_i+\alpha s_i) < J(u_i) \tag{3.5}$$

Thus for all $\alpha > 0$

$$\frac{J(u_i+\alpha s_i) - J(u_i)}{\alpha} \geq 0 \tag{3.6}$$

In the limit as α approaches zero this gives

$$(g_i, s_i) = -\,||g_i||^2 \geq 0 \tag{3.7}$$

which contradicts the original assumption and hence there exists $\alpha > 0$ such that $J(u_i+\alpha s_i) < J(u_i)$. Since α_i is chosen to minimize $J(u_i+\alpha s_i)$ the theorem follows.

The sequence of real numbers $\{J(u_i)\}$ is thus monotone decreasing and therefore has a limit J_∞ in the extended real numbers.

Also of interest is the limiting behavior of the sequences $\{u_k\}$ and $\{g_k\}$. Results similar to those that have been obtained for the method of steepest descent [see ref. 12,13] are given below.

Theorem 3

If 1. $J(u)$ is bounded below
2. $J(u)$ and $g(u)$ are continuous

3. $D^2J(u,h,h)$ exists and
$|D^2 J(u,h,h)| \leq m \,||h||^2 \quad \forall u, h \in H$ and $m > o$
4. $\{u_k\}$ has a cluster point , u^*

then the sequence u_k formed with arbitrary u_0 by the conjugate gradient method has the following properties

1. $\lim\limits_{k\to\infty} J(u_k) = J(u^*)$
2. $\lim\limits_{k\to\infty} g(u_k) = g(u^*) = 0$

The proof of this theorem is not presented here due to lack of space.

4. LINEAR QUADRATIC CASE

In order for a minimization technique to perform efficiently on a general function it is necessary, as has been emphasized previously [see ref. 5,6] , that it efficiently minimize a quadratic function. This follows since near the minimum, the quadratic terms in the Taylor series expansion predominate. For the problem of (2.1) - (2.3) this corresponds to the linearization of (2.2) about the minimum and the approximation of (2.1) by a positive definite quadratic form, i.e., the Hilbert space problem becomes the minimization of the quadratic function,

$$J(u) = \frac{1}{2} (u-u^*, A(u-u^*)) \tag{4.1}$$

where A is a positive definite self adjoint linear operator. Then, the following results are shown by Antosiewicz and Rheinboldt in Todd see ref. 11 :

1. There exists a unique $u = u^*$ which minimizes $J(u)$ on H and for which $g(u^*) = 0$. In addition, it is shown that the conjugate gradient method, when applied to this function, has the following properties.
2. The set $\{s_i\}$ satisfies $(s_i, As_j) = 0 \quad i \neq j$. This implies that $\{s_i\}$ is a linearly independent set.
3. At the n^{th} iteration $J(u)$ is minimized over the set $B_n = \{u|u=u_0 + \sum_{j=0}^{n-1} \alpha_j s_j\}$. Since the s_i are linearly independent $B_{j-1} \subset B_j \quad j = 1,\ldots,n$ and hence $J(u)$ is minimized over an expanding sequence of sets.
4. The error vector $y_i = u^* - u_i$ is decreased at each step, i.e., $||u^* - u_i||<||u^* - u_{i-1}||$ if $g(u_{i-1}) \neq 0$.
5. If the sequence of sets $\{B_n\}$ covers H , that is $\bigcup_{n=1}^{\infty} B_n = H$ then $u_i \to u^*$ as $i \to \infty$

All the above properties, with the exception of 4, hold true for any method which generates conjugate directions when applied to a quadratic function. None of these properties are shared by the method of steepest descent.

In the (finite) n-dimensional case, property 3 implies that the quadratic function will be minimized by any conjugate direction method in at most n steps. This is in contrast to the method of steepest descent, for which it has recently been shown that finite convergence is not obtained unless it is obtained on the first step. [see ref. 14]

4.1 Further Results

The question of relative performance of the steepest descent and conjugate gradient methods on linear-quadratic problems is of much interest. A number of results follow from the fact that the conjugate gradient method minimizes $J(u)$ over an expanding sequence of sets, B_i .

Let barred and unbarred quantities be those achieved by the conjugate gradient and steepest descent methods respectively. By (2.15) the sets B_i of result 3 of this section are also given by

$$B_i = \{u | u = \bar{u}_0 + \sum_{j=0}^{i-1} \alpha_j \bar{g}_j\} \tag{4.2}$$

Since $u_0 = \bar{u}_0$ and $u_1 = \bar{u}_1$ then $g_1 = \bar{g}_1$ and both u_2 and $\bar{u}_2$ are of the form $u_0 + \alpha_0 g_0 + \alpha_1 g_1$. Since the conjugate gradient method chooses the α_i to minimize $J(u)$, then

$$J(\bar{u}_2) \leq J(u_2) \tag{4.3}$$

Similar reasoning applies to the situation where, at step i of the conjugate gradient method, one reverts to the method of steepest descent by setting $\beta_{i-1} = 0$ in (2.15). The value of J obtained at the next step is never smaller than the value obtained if the conjugate algorithm has been continued.

A more difficult question is whether one method always does better than the other at every iteration, given the same starting point for both. This can be resolved for the case of one state variable and infinite final time as follows. Consider the problem

$$\text{minimize} \qquad J(u) = \frac{1}{2} \int_0^\infty (x^2+ku^2)dt \tag{4.4}$$

$$\text{subject to} \qquad \dot{x} = ax + bu, \quad x(0) = c \tag{4.5}$$

where all quantities are scalar functions of time.

The relations defining the gradient, (2.5) - (2.7) become

$$x = ax + bu \qquad x(0) = c, \quad a < 0 \tag{4.6}$$

$$\lambda = a\lambda - x \qquad \lambda(\infty) = 0 \tag{4.7}$$

$$g(u) = \frac{\partial H}{\partial u} = b\,\lambda + ku \tag{4.8}$$

Assume $u_0 = 0$. (If not, a shift or origin reproduces (4.6) - (4.8)). Equations (4.6) and (4.7) then both have solutions of the form $c_0 e^{at}$, since the conditions $\lambda(\infty) = 0$ and $a < 0$ allow no term in e^{-at} in the solution of (4.7). Thus both $g(u_0)$ and u_1 are of this form. Using this u_1 in (4.6) yields a solution of the form $e^{at}(r_0+r_1t)$ for (4.6) and (4.7) and $g(u_1)$, u_2 and $\bar{u}_2$ have this form also. A simple inductive argument shows that g_n, $\bar{g}_n$, u_n, and $\bar{u}_n$ are all of the form

$$e^{at} P_{n-1}(t)$$

where P_{n-1} is a polynomial of degree $n - 1$ in t . Thus both u_n and

$\bar{u}_n$ are elements of the set B_n in (4.2) and, since $\bar{u}_n$ minimizes $J(u)$ over B_n

$$J(\bar{u}_n) \leq J(u_n), \qquad n = 0, 1, \ldots \tag{4.9}$$

5. COMPUTATIONAL RESULTS

5.1 Introduction

In order to compare the performance of this method with that of steepest descent, three numerical examples are given. The steepest descent technique used is obtained from the conjugate gradient algorithm by setting $\beta_i = 0$ in (2.15). The linear search was performed using a cubic interpolation technique [see ref. 6] and is the same for both methods. Thus both methods require about the same computational effort per iteration. Comparison with a second variational method is also given in example 3.

5.2 Examples

Example 1. This first example considers the linear system

$$\dot{x}_1 = x_2, \quad x_1(0) = 1 \tag{5.1}$$

$$\dot{x}_2 = -x_2 + u, \quad x_2(0) = -1 \tag{5.2}$$

and quadratic functional

$$J(u) = \int_0^1 \{(x_1-1)^2 + x_2^2 + 0.005\,u^2\}dt \tag{5.3}$$

This is the problem considered by Hsieh see ref. 15 .

Figure 1 shows the variation of the function (5.3) with the number of iterations for the steepest descent and the conjugate gradient methods. The conjugate gradient technique has effectively converged to the optimum in four iterations, whereas the steepest descent scheme shows slow convergence near the optimum, requiring fifteen iterations to achieve similar results.

Example 2. This example is the rocket launch problem described by Dreyful [see ref. 16].

$$\dot{x}_1 = x_2 \tag{5.4}$$

$$\dot{x}_2 = 64 \sin u - 32 \tag{5.5}$$

$$\dot{x}_3 = 64 \cos u \tag{5.6}$$

$$x_i(0) = 0 \quad i = 1, 2, 3 \tag{5.7}$$

with objective function

$$\max x_3(100) \tag{5.8}$$

subject to

$$x_1(100) = 100,000 \tag{5.9}$$

$$x_2(100) = 0 \tag{5.10}$$

Using a penalty function this problem was converted to a problem with no end point constraints;

$$\min -x_3(100) + 0.002(x_1(100) - 105)^2 + 0.05(x_2(100))^2 \tag{5.11}$$

subject to (5.6) - (5.9). Figure 2 shows the convergence of the objective function (5.11).

Example 3. As a final example, and one which is nonlinear in the state variables, consider the problem suggested by Merriam [see ref. 17].

$$\dot{x}_1 = (1-x_2^2)\,x_1 - x_2 + u, \quad x_1(0) = 0 \tag{5.12}$$

$$\dot{x}_2 = x_1, \quad x_2(0) = 3 \tag{5.13}$$

$$\dot{x}_3 = x_1^2 + x_2^2 + u^2, \quad x_3(0) = 0 \tag{5.14}$$

with objective

$$\min x_3(10) \tag{5.15}$$

Figure 3 compares the convergence of the steepest descent, conjugate gradients, and the second variational method used by Merriam. As in the previous examples the conjugate gradient method is markedly superior to steepest descents. The second variational technique is faster than both but requires considerably more computation per iteration. Figure 4 compares the gradient trajectories for steepest descent and conjugate gradients.

7. CONCLUSION

The conjugate gradient method presented here appears to have significant advantages over existing methods. Its convergence rate is evidently superior to that of steepest descents with very little additional computation per iteration. Since the method always generates directions of descent, its stability properties are superior to those of second variational techniques, which may not converge from a poor approximation to the minimum. Second variational methods will probably have superior local convergence, but at greater computational cost. Thus the conjugate gradient methods seems to be an effective compromise between existing techniques.

It is apparent that many of the other existing finite dimensional minimization techniques can be similarly extended. For problems with inequality constraints, an extension of the SUMT procedure of Fiacco and McCormick [see ref. 18] appears promising. In particular, combining SUMT with conjugate gradients should yield a computationally useful algorithm. Such an extension is currently being investigated.

REFERENCES

1. R. E. Kopp and R. McGill, "Several Trajectory Optimization Techniques," Computing Methods in Optimization Problems, edited by A.V. Balakrishnan and L.W. Nenstadt, pp. 65-89, Academic Press, New York, 1964.

2. H. J. Kelley, "Methods of Gradients," Chapter 6 of Optimization Techniques, edited by G. Leitmann, Academic Press, New York, 1962.

3. W. F. Denham and A. E. Bryson, "Optimal Programming Problems with Inequality Constraints II: Solution by Steepest-Ascent," AIAA Journal, Vol. 2, No. 1, pp. 25-34, January 1964.

4. R. McGill, "Optimal Control, Inequality State Constraints and the Generalized Newton-Raphson Algorithm," J. SIAM Control, Series A, Vol. 3, No. 2, pp. 291-298, 1965.

5. R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," British Computer Journal, pp. 163-168, June 1963.

6. R. Fletcher and C. M. Reeves, "Function Minimization by Conjugate Gradients," British Computer Journal, pp. 149-154, July 1964.

7. M. R. Hestenes and E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Systems," J. Res. N.B.S., Vol. 49, p 409, 1952.

8. H. J. Kelley, "Gradient Theory of Optimal Flight Paths," ARS Journal, Vol. 30, pp. 947-953, October 1960.

9. A. E. Bryson and S. R. McReynolds, "A Successive Sweep Method for Solving Optimal Programming Problems," 1965 JACC Proceedings, Troy, N.Y.

10. S. Mitter, "Successive Approximation Methods for the Solution of Optimal Control Problems," Automatica, Vol. 3, pp. 133-149, 1966.

11. J. Todd, Editor, Survey of Numberical Analysis, McGraw Hill Book Company, New York, N. Y., 1962.

12. J. R. Rice, The Approximation of Functions, Vol. 1, Addison Wesley Publishing Co., Reading, Mass., 1964.

13. H. B. Curry, "The Method of Steepest Descent for Nonlinear Minimization Problems," Quarterly of Applied Mathematics, Vol. 2, No. 3, pp. 258-261, October, 1944.

14. W. E. Langlois, "Conditions for Termination of the Method of Steepest Descent After a Finite Number of Iterations," I.B.M. J. of Res. and Dev., pp. 98-99, January 1966.

15. C. T. Leondes, editor, Advances in Control, Vol. II.

16. S. E. Dreyfus, "Variational Problems with State Variable Inequaltiy Constraints," Rand Report P-2605-1, August 1963.

17. C. W. Merriam, III, "Direct Computational Methods for Feedback Control Optimaization," Information and Control, April 1965.

18. A. V. Fiacco and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, a Primal-Dual Method," Management Science, Vol. 10, No. 2, pp. 360-366, January 1964.

FIGURE 1

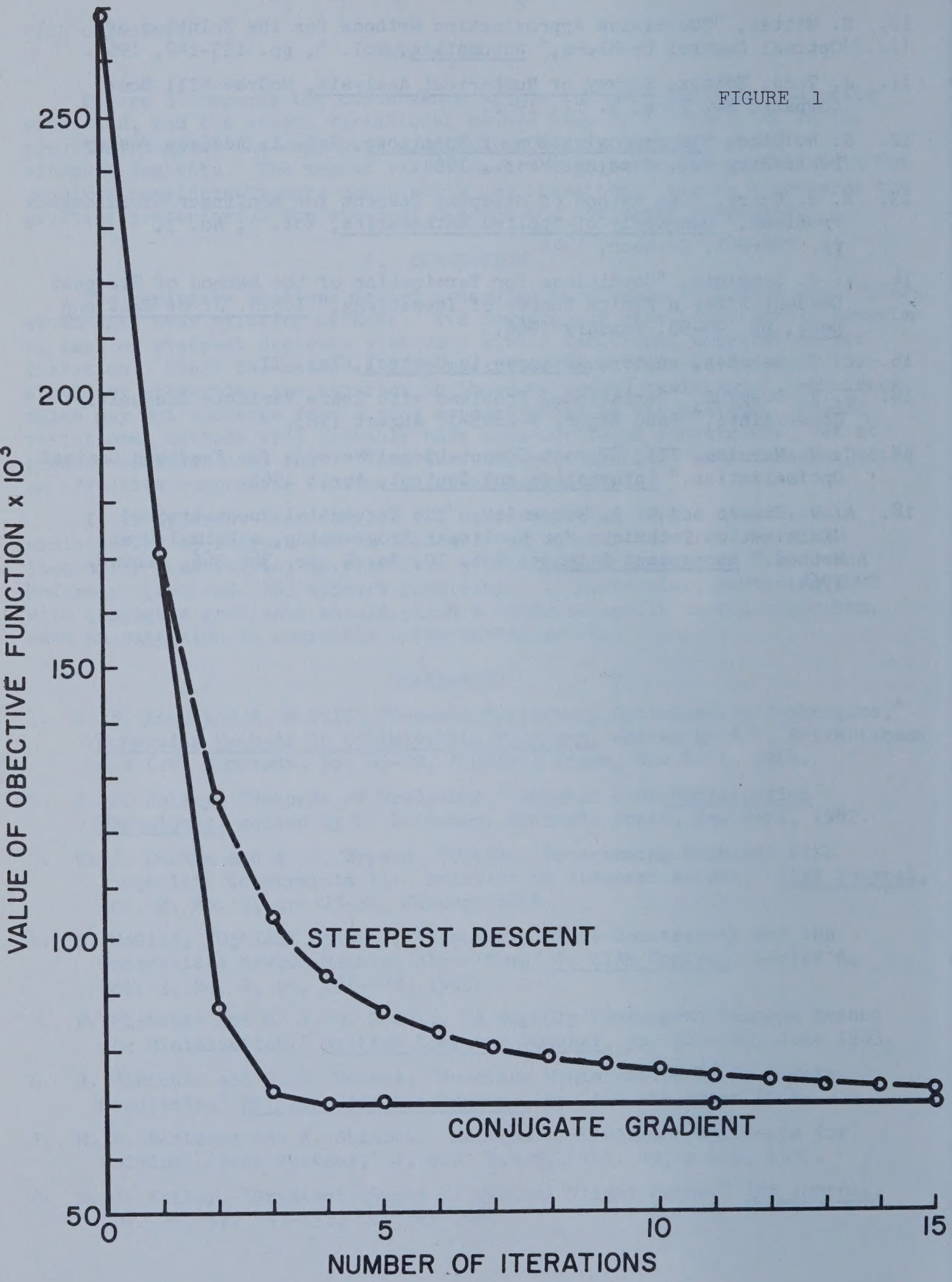
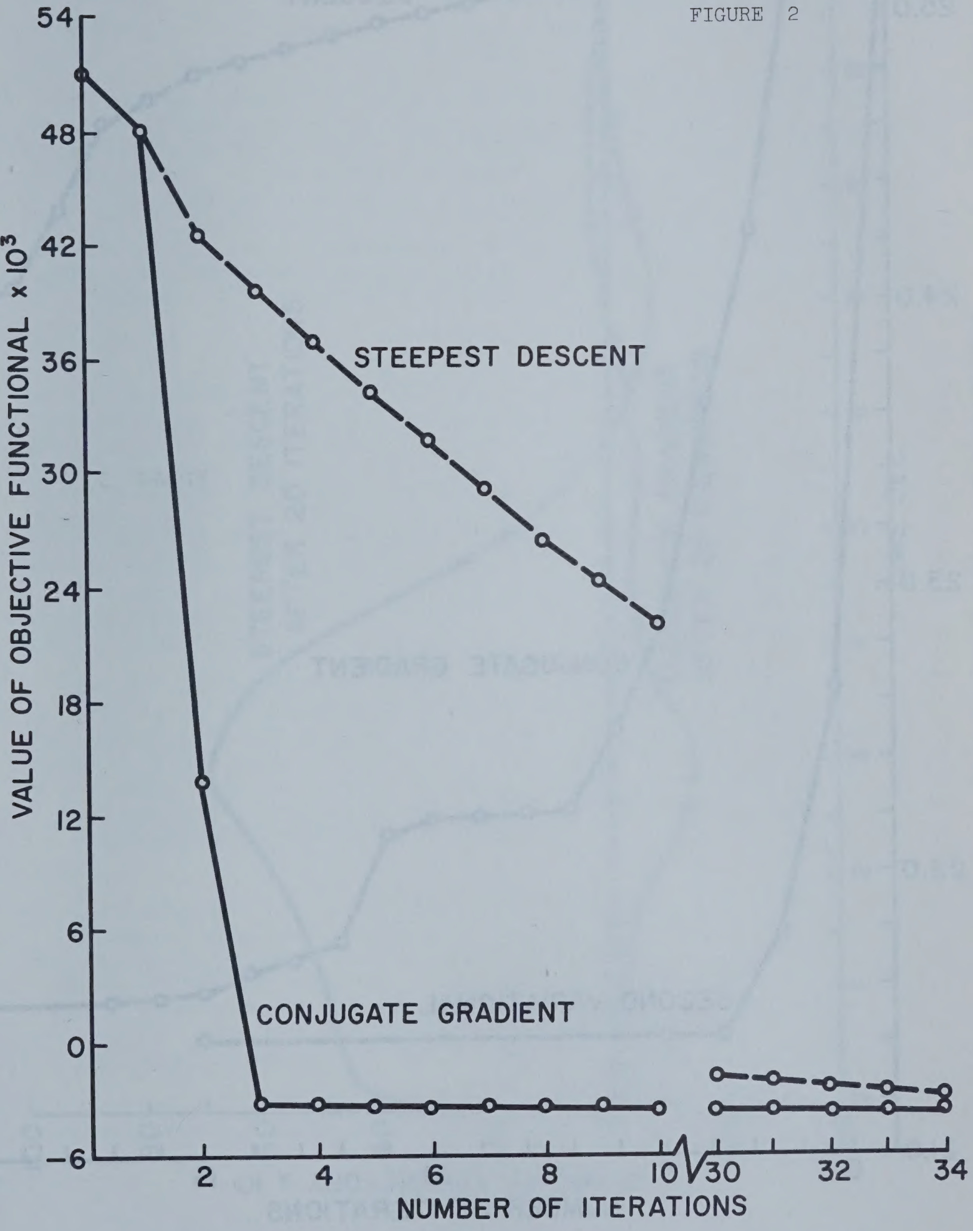

VALUE OF OBECTIVE FUNCTION x 10^3
250
200
150
100
50
0
5
10
15
NUMBER OF ITERATIONS
STEEPEST DESCENT
CONJUGATE GRADIENT

FIGURE 2


VALUE OF OBJECTIVE FUNCTIONAL x 10^3
54
48
42
36
30
24
18
12
6
0
−6
2
4
6
8
10
30
32
34
NUMBER OF ITERATIONS
STEEPEST DESCENT
CONJUGATE GRADIENT

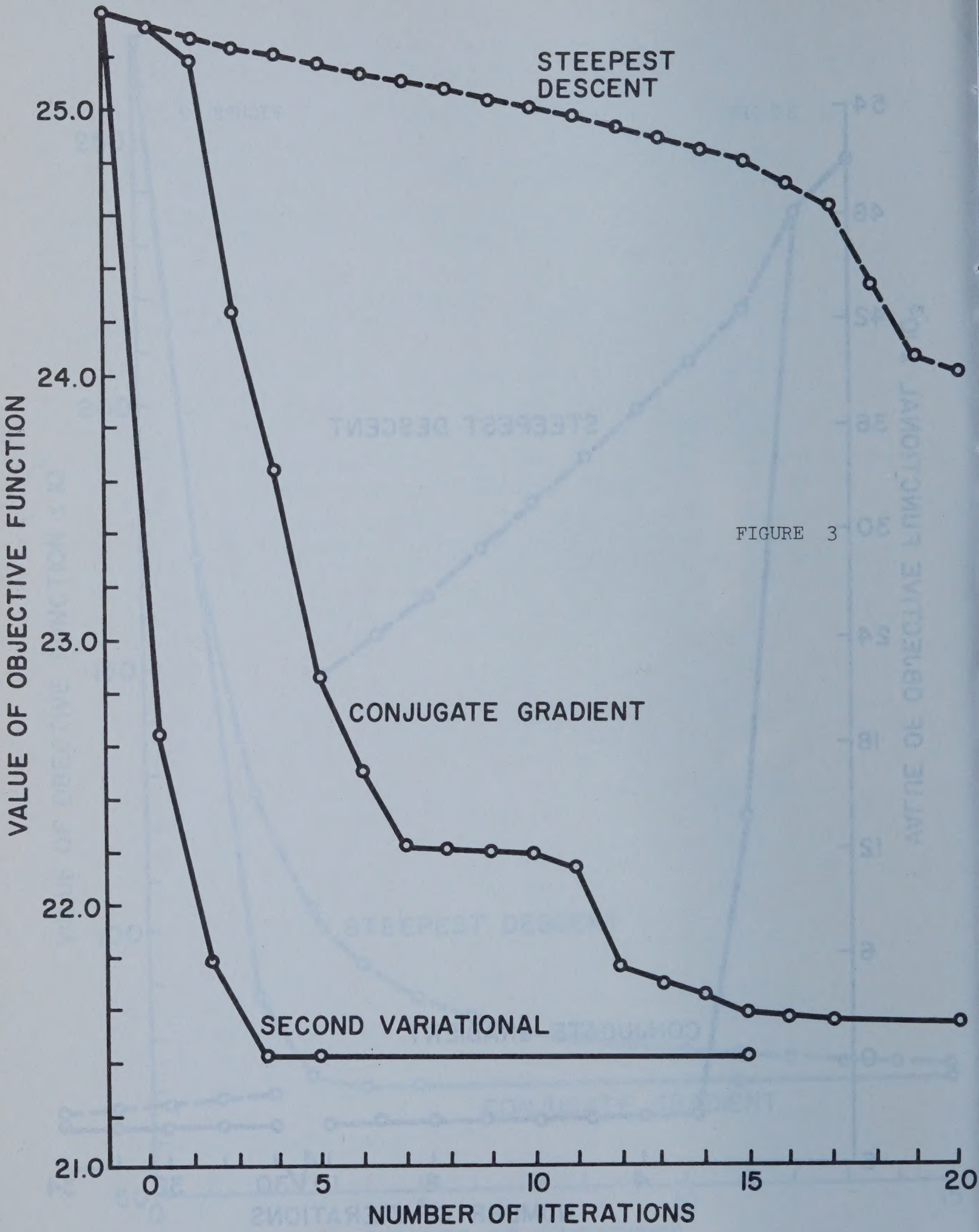
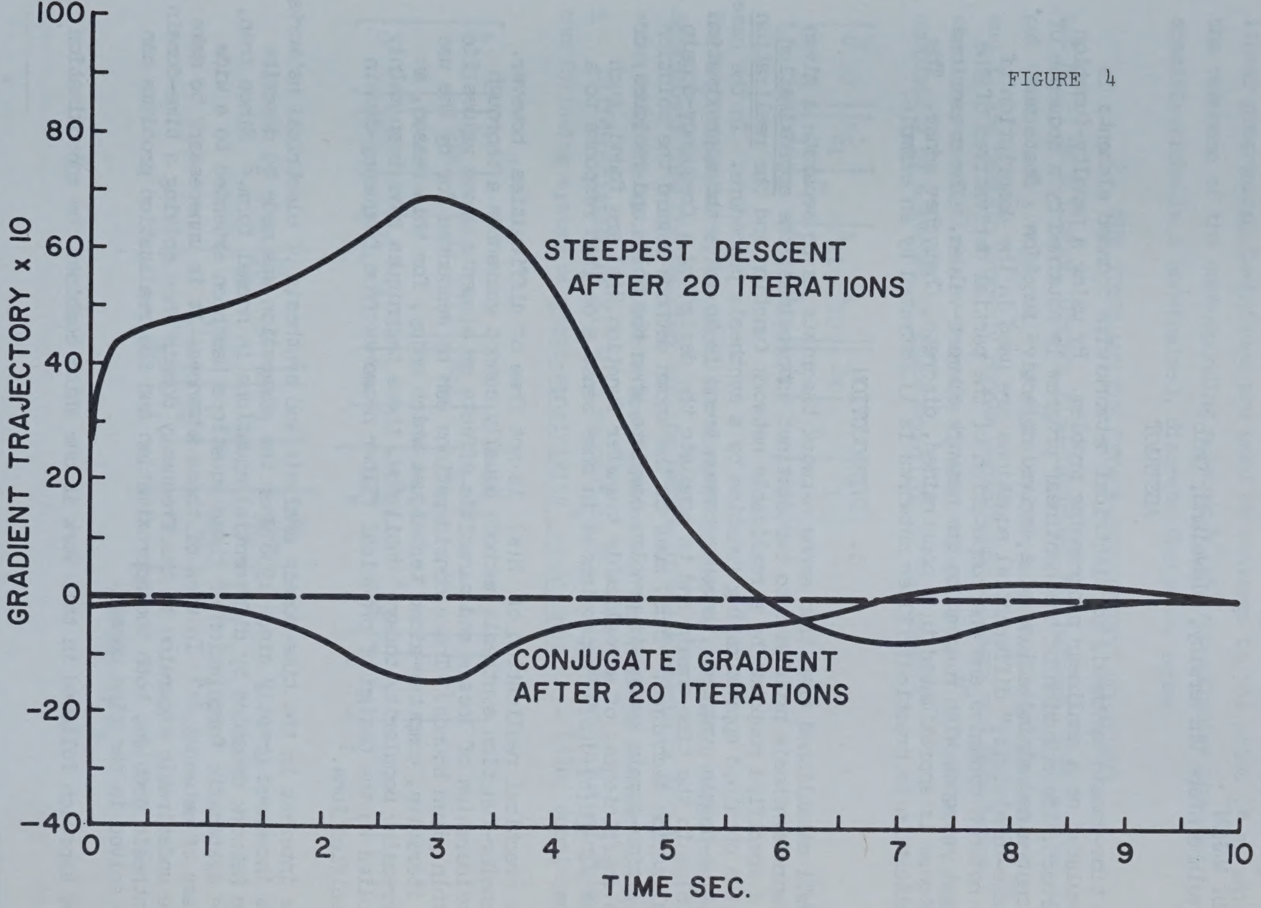

STEEPEST DESCENT
CONJUGATE GRADIENT
SECOND VARIATIONAL
VALUE OF OBJECTIVE FUNCTION
25.0
24.0
23.0
22.0
21.0
0
5
10
15
20
NUMBER OF ITERATIONS

FIGURE 3

FIGURE 4


GRADIENT TRAJECTORY x 10
100
80
60
40
20
0
-20
-40
STEEPEST DESCENT
AFTER 20 ITERATIONS
CONJUGATE GRADIENT
AFTER 20 ITERATIONS
0
1
2
3
4
5
6
7
8
9
10
TIME SEC.

TIME-DOMAIN SYNTHESIS BY NONLINEAR PROGRAMMING[*]

I. A. ATHANASSOPOULOS, J. D. SCHOEFFLER
Case Institute of Technology, Cleveland, Ohio

A. D. WAREN
Cleveland State University, Cleveland, Ohio

ABSTRACT

The time-domain synthesis of electrical networks with bounded elements is formulated as a nonlinear programming problem. By using a penalty-function approach, the solution of the nonlinear program is obtained by a sequence of unconstrained minimizations of a derived objective function. State and "state-sensitivity" differential equations are used in the description of the network dynamics and the computation of the partial derivatives of the output response with respect to the network element-values. The prescribed response is approximated in a constrained, discrete, Chebychev sense. The application to practical filter networks is illustrated by an example.

1. INTRODUCTION

A well established tradition among network theorists is to separate a given network synthesis problem into two distinct subproblems: the approximation of a specified response by a realizable network function and the realization of the obtained approximating function by a synthesis procedure. In the case of time-domain synthesis, another common trend is to solve the approximation problem in the time-domain and to complete the design by a frequency-domain realization technique. Thus, most of the known methods regard the solution of a time-domain synthesis problem complete when the poles and residues, or the coefficients, of a realizable transfer function H(s) are found, such that $\mathcal{L}^{-1}[H(s)\cdot U(s)]$ approximates in some sense a desired response to a given input $u(t)$.

The practical realization of $H(s)$ is not free of difficulties, however. Transfer-function synthesis methods usually cannot guarantee a thorough consideration of losses and parasitic effects or elements whose values lie within given bounds. These considerations can be accounted for by the use of iterative, computer-aided techniques which enjoy, for this reason, an increasing popularity today. Until now, these techniques have been mainly applied to the design of practical filter networks from frequency-domain specifications.[1]

The interest in the time-domain analysis and synthesis of electrical networks has increased greatly since 1958 when the suggestion was made to describe the network dynamics by differential equations in normal form.[2] Since then, the systematic formulation of these equations has been extended to a wide class of networks.[3,4] In view of these advances, it is unnecessary to make the undesirable excursion to the frequency domain when solving a time-domain synthesis problem; both the approximation and the realization problems can be solved in the time domain.

The approach followed in this work is one which combines the approximation

---

[*] This work was supported by the Air Force Cambridge Research Laboratories, Office of Aerospace Research under Contract AF 19(628) 1699.

and realization problems by considering a network $\mathcal{N}(\underline{\varepsilon})$ of N_e elements $\underline{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_{Ne}]^T$ and fixed topology as a point in an N_e- dimensional Euclidean space. Starting from an initial network $\mathcal{N}(\underline{\varepsilon}_o)$, efficient nonlinear programming techniques are used to converge to the point $\underline{\varepsilon}^*$ at which the response of the corresponding network $\mathcal{N}(\underline{\varepsilon}^*)$ approximates the given specifications in a constrained, discrete Chebychev sense.

2. THE STATE AND THE "STATE-SENSITIVITY" EQUATIONS

Let $\mathcal{N}(\underline{\varepsilon})$ be a linear, lumped, time-invariant RLC network, $\underline{u}(t)$ a given input vector and $\underline{x} = [\underline{x}_1, \underline{x}_2]^T$ a state vector, such that the elements of $\underline{x}_1$ and $\underline{x}_2$ are certain properly chosen capacitance voltages and inductance currents, respectively. Then the dynamic behavior of $\mathcal{N}(\underline{\varepsilon})$ can be described completely by a system of state equations of the form:[3,4]

$$\begin{bmatrix} \tilde{C} & 0 \\ 0 & \tilde{L} \end{bmatrix} \begin{bmatrix} \dot{\underline{x}}_1 \\ \dot{\underline{x}}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \underline{u} \qquad (1)$$

For a given network, the explicit form of the above matrices, which can be determined from the topology and the nature of the elements of $\mathcal{N}(\underline{\varepsilon})$[3,4], shows that the matrix elements are functions of the network element values $\underline{\varepsilon}$. Differentiation of (1) with respect to the k^{th} element value ε_k gives the following system of state-sensitivity* equations:

$$\begin{bmatrix} \tilde{C} & 0 \\ 0 & \tilde{L} \end{bmatrix} \begin{bmatrix} \dot{\underline{S}}_{1k} \\ \dot{\underline{S}}_{2k} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \underline{S}_{1k} \\ \underline{S}_{2k} \end{bmatrix} + \begin{bmatrix} \frac{\partial A_{11}}{\partial \varepsilon_k} & \frac{\partial A_{12}}{\partial \varepsilon_k} \\ \frac{\partial A_{21}}{\partial \varepsilon_k} & \frac{\partial A_{22}}{\partial \varepsilon_k} \end{bmatrix} \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix}$$

$$- \begin{bmatrix} \frac{\partial \tilde{C}}{\partial \varepsilon_k} & 0 \\ 0 & \frac{\partial \tilde{L}}{\partial \varepsilon_k} \end{bmatrix} \begin{bmatrix} \dot{\underline{x}}_1 \\ \dot{\underline{x}}_2 \end{bmatrix} + \begin{bmatrix} \frac{\partial B_1}{\partial \varepsilon_k} \\ \frac{\partial B_2}{\partial \varepsilon_k} \end{bmatrix} \underline{u} \qquad (2)$$

$$k = 1, \ldots, N_e$$

---

*Note that the meaning of sensitivity in this section differs from the usual definition of this term.

where, by definition,

$$\underline{S}_{1k} = \frac{\partial \underline{x}_1}{\partial \varepsilon_k} \quad , \quad \underline{S}_{2k} = \frac{\partial \underline{x}_2}{\partial \varepsilon_k}$$

The network output $\underline{y}$ is obtained from the output equations

$$\underline{y} = C\,\underline{x} + D\,\underline{u} \tag{3}$$

and the derivatives of $\underline{y}$ with respect to ε_k are given by the output-sensitivity equations

$$\frac{\partial \underline{y}}{\partial \varepsilon_k} = [C]\underline{S}_k + \left[\frac{\partial C}{\partial \varepsilon_k}\right]\underline{x} + \left[\frac{\partial D}{\partial \varepsilon_k}\right]\underline{u} \qquad k = 1,\ldots, N_e \tag{4}$$

where $\underline{S}_k \equiv [\underline{S}_{1k} \, , \, \underline{S}_{2k}]^T$.

3. CONSTRAINED DISCRETE CHEBYCHEV APPROXIMATION

Let the time axis be regarded as a sequence of equidistant points $t_i = ih$ and $\{y_j^*(t_i) \mid \; i = 1,\ldots,N_T\}$ be an ordered set of specified values for the j^{th} component of the output vector $\underline{y}(t)$ of a network $\mathcal{N}(\underline{\varepsilon})$ at the first N_T points. Then the problem of discrete Chebychev approximation[5] is to find the element vector $\varepsilon = \varepsilon^*$ for which the response $y_j(\underline{\varepsilon}^*;t)$ of $\mathcal{N}(\underline{\varepsilon}^*)$ is such that

$$E_\infty \equiv \max_{0 \le i \le N_T} \left| y_j^*(t_i) - y_j(\underline{\varepsilon}^*;t_i) \right| \tag{5}$$

is minimum over all $\underline{\varepsilon}$.

The realizability of $\mathcal{N}(\underline{\varepsilon}^*)$ requires that $\underline{\varepsilon}^* \geq \underline{0}$. In most practical design problems, however, the range of acceptable $\underline{\varepsilon}^*$ is further constrained:

$$\underline{L} \leq \underline{\varepsilon}^* \leq \underline{U} \; ; \; \underline{L} \geq \underline{0} \tag{6}$$

so that the designed network can be constructed by practically available components. The presence of the constraints (6) will be recognized by referring to the problem of minimizing (5) subject to (6) as a constrained, discrete Chebychev approximation problem. Approximations of this type are very useful in time-domain network synthesis because they can allow direct control of quantities such as the maximum over- or undershoot and the delay time. This error criterion (without element-value constraints) has also been used in the approximation of a specified impulse response by realizable network functions[6,7,8,9] .

In this work, the constrained discrete Chebychev approximation problem (5) and (6) is formulated as a problem of nonlinear programming. Refer to Fig. 1, introduce an additional independent variable η and write the following approximation inequalities:

$$\psi_1(\underline{\varepsilon},\eta;t_i) \equiv y_j^*(t_i) + w_{ij}\,\eta - y_j(\underline{\varepsilon};t_i) \geq 0 \tag{7}$$

$$\psi_2(\underline{\varepsilon},\eta;t_i) \equiv y_j(\underline{\varepsilon},t_i) + w_{ij}\,\eta - y_j^*(t_i) \geq 0 \tag{8}$$

Here w_{ij} are positive weights which may be used to extend the flexibility of the approximation. If all $w_{ij} = 1$, then the following nonlinear program is equivalent to the problem of minimizing (5) subject to (6):

$$\underset{\underline{\varepsilon},\,\eta}{\text{minimize}} \quad \eta \tag{9a}$$

subject to

$$\psi_1(\underline{\varepsilon},\eta;t_i) \geq 0 \qquad 1 \leq i \leq N_T \tag{9b}$$

$$\psi_2(\underline{\varepsilon},\eta;t_i) \geq 0 \qquad 1 \leq i \leq N_T \tag{9c}$$

$$\varepsilon_k - L_k \geq 0 \qquad 1 \leq k \leq N_e \tag{9d}$$

$$U_k - \varepsilon_k \geq 0 \qquad 1 \leq k \leq N_e \tag{9e}$$

An initial feasible solution is easily found by selecting η sufficiently large.

4. PENALTY FUNCTION FORMULATION

When, as in the case of the approximation inequalities (7) and (8), the constraints are highly nonlinear functions of the problem variables, the solution of a nonlinear program of the form of (9) is efficiently accomplished through a conversion to an unconstrained optimization problem by means of a penalty function formulation[10]. In particular, this work follows a recent technique of Fiacco and McCormick[11] which shows that the solution of (9) may be obtained by a sequence of unconstrained minimizations of the function

$$P(\underline{\varepsilon},\eta;r_\mu) = \eta + r_\mu \sum_{i=1}^{N_T} \left[\frac{1}{\psi_1(\underline{\varepsilon},\eta;t_i)} + \frac{1}{\psi_2(\underline{\varepsilon},\eta;t_i)}\right] + r_\mu \sum_{k=1}^{N_e} \left[\frac{1}{\varepsilon_k - L_k} + \frac{1}{U_k - \varepsilon_k}\right] \tag{10}$$

for a strictly monotonic decreasing sequence of positive numbers $< r_\mu >$. This method has also been applied to several other computer-aided design problems.[1,12]

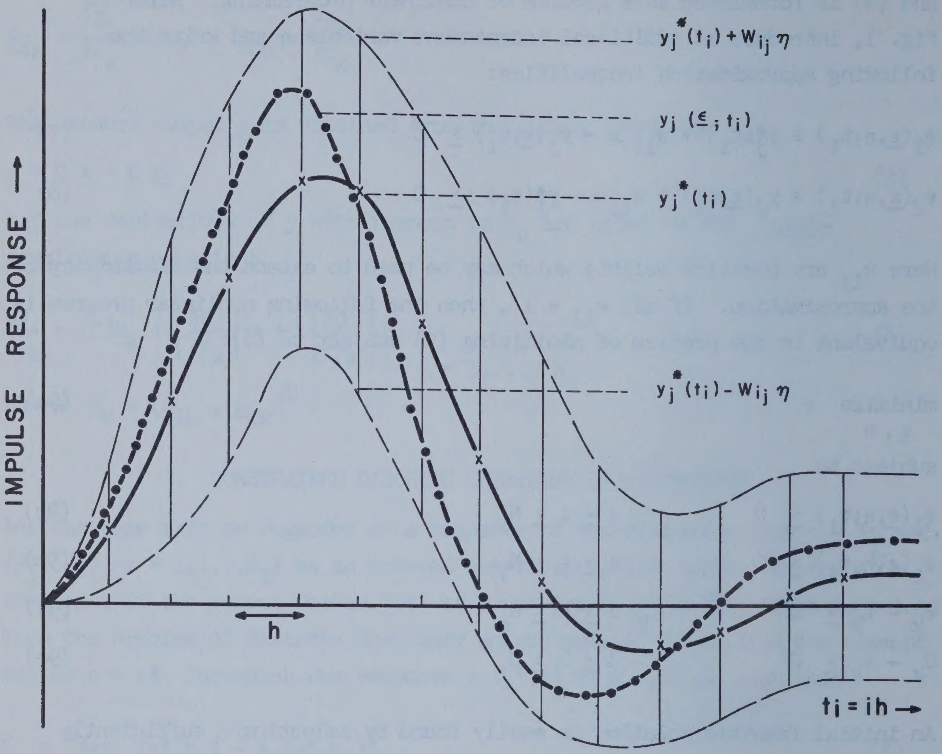


FIG. 1 Approximation of a specified time-domain response

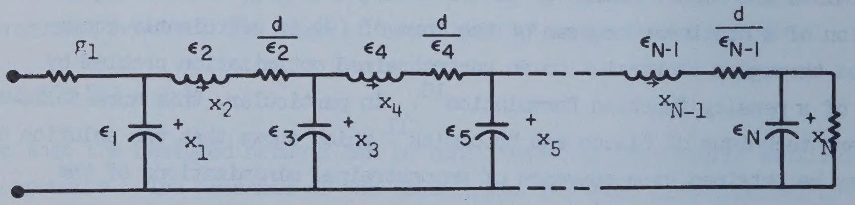

g_1
ϵ_2
$\frac{d}{\epsilon_2}$
ϵ_4
$\frac{d}{\epsilon_4}$
ϵ_{N-1}
$\frac{d}{\epsilon_{N-1}}$
x_2
x_4
x_{N-1}
ϵ_1
$+ x_1$
ϵ_3
$+ x_3$
ϵ_5
$+ x_5$
ϵ_N
$+ x_N$
g_2

FIG. 2 A low-pass ladder network

5. THE UNCONSTRAINED MINIMIZATION OF $P(\underline{\varepsilon},\eta;r_\mu)$

A remaining problem is that of finding the successive minima of the P-function (10) in the (N_e+1)-dimensional space of $\hat{\underline{\varepsilon}} \equiv [\varepsilon,\eta]^T$. Of the various methods for such a task, a number of highly efficient techniques make use of the first derivatives of the function. Since the derivatives of $P(\hat{\underline{\varepsilon}};r_\mu)$ are easily computed,

$$\frac{\partial P(\hat{\underline{\varepsilon}};r_\mu)}{\partial \varepsilon_k} = r_\mu \sum_{i=1}^{N_T} S_{kj}(t_i)[\psi_1^{-2}(\hat{\underline{\varepsilon}};t_i)-\psi_2^{-2}(\hat{\underline{\varepsilon}};t_i)]+r_\mu[(U_k-\varepsilon_k)^{-2}-(\varepsilon_k-L_k)^{-2}] \qquad (11)$$

$$S_{kj}(t_i) \equiv \partial y_j(t_i)/\partial \varepsilon_k \quad ; \quad k = 1,\ldots,N_e$$

$$\frac{\partial P(\hat{\underline{\varepsilon}};r_\mu)}{\partial \eta} = 1 + r_\mu \sum_{i=1}^{N_T} w_{ij}[\psi_1^{-2}(\hat{\underline{\varepsilon}};t_i)-\psi_2^{-1}(\hat{\underline{\varepsilon}};t_i)] \ , \qquad (12)$$

the use of gradient methods is possible here.

The unconstrained minimization method used in this work is a gradient technique developed by Fletcher and Powell.[13] Its adoption is supported by the results of a recent systematic comparison of several optimization methods[14] and by the success of an earlier application of this technique to network synthesis from frequency-domain specifications.[1]

Let H_o be any $(N_e+1)\mathrm{x}(N_e+1)$ positive definite matrix. Then an iteration of the Fletcher-Powell algorithm is outlined as follows:

1. Find the direction $\underline{s}_i = -H_i \ \nabla P(\hat{\underline{\varepsilon}}_i;r_\mu)$

   where $\nabla P(\hat{\underline{\varepsilon}}_i;r_\mu) = [\frac{\partial P}{\partial \varepsilon_1}, \ \ldots \ , \frac{\partial P}{\partial \varepsilon_{N_e}}, \frac{\partial P}{\partial \eta}]^T$.

2. Minimize P along $\underline{s}_i$, i.e. find $\alpha_i = \alpha_i^*$ by minimizing $P(\hat{\underline{\varepsilon}}+\alpha\underline{s}_i;r_\mu)$.

3. Update $\hat{\underline{\varepsilon}}_i$ and H_i :

   $\underline{\sigma}_i = \alpha_i^* \ \underline{s}_i \quad , \quad \hat{\underline{\varepsilon}}_i = \hat{\underline{\varepsilon}}_i + \underline{\sigma}_i$

   $H_{i+1} = H_i + A_i + B_i$ where

   $$A_i = \frac{\underline{\sigma}_i \ \underline{\sigma}_i^T}{\underline{\sigma}_i^T \ \underline{\delta}_i} \quad ; \quad \underline{\delta}_i \equiv \nabla P(\hat{\underline{\varepsilon}}_{i+1};r_\mu) - \nabla P(\hat{\underline{\varepsilon}}_i;r_\mu)$$

   $$B_i = \frac{-H_i \ \underline{\delta}_i \ \underline{\delta}_i^T \ H_i}{\underline{\delta}_i^T \ H_i \ \underline{\delta}_i}$$

The efficiency of this method depends critically on the efficiency of the one-dimensional minimization of step 2 above. To avoid the several gradient computations of higher-order interpolation schemes and the slow convergence of techniques requiring only functional evaluations, a quadratic interpolation method was used, requiring one gradient and a few function evaluations per step. It should be noted here that each evaluation of $P(\hat{\underline{\varepsilon}};r_k)$ requires the (numerical) solution of the state equations (1) while the computation of $\nabla P(\hat{\underline{\varepsilon}};r_k)$ must follow the solution of N_e sets of state-sensitivity equations.

6. EXAMPLE

Let $\mathcal{N}(\underline{\varepsilon})$ be the low pass ladder network of Fig. 2, described by the following system of state equations

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \vdots \\ \dot{x}_N \end{bmatrix} = \begin{bmatrix} -\varepsilon_1 g_1 & -\varepsilon_1 & & & & \\ \varepsilon_2 & -d & -\varepsilon_2 & & & \\ & \varepsilon_3 & 0 & -{}_3 & & \\ & & \varepsilon_4 & -d & -\varepsilon_4 & \\ & & & & & \\ & & & & \varepsilon_N & -\varepsilon_N g_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 g_1 e(t) \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \qquad (13)$$

where $\varepsilon_1 \equiv 1/C_1$, $\varepsilon_2 \equiv 1/L_2$,... etc. and $d = R/L$ is a uniform dissipation factor. The state-sensitivity equations are derived by differentiation:

$$\dot{\underline{s}}_k = A\underline{s}_k + A'_k\,\underline{x} + B'_k\,\underline{u} \quad ; \quad k = 1,\dots,N \qquad (14)$$

where A is as in (13) and

$$A'_1 = \left[\begin{array}{cc:c} -g_1 & -1 & \\ \hdashline & & \\ & & \mathbf{O} \end{array}\right] ; \quad A'_N = \left[\begin{array}{c:cc} \mathbf{O} & & \\ & & \\ \hdashline & 1 & -g_2 \end{array}\right]$$

$$A'_k = \begin{array}{c} \quad\quad\quad\quad k^{th}\ \text{column} \\ \left[\begin{array}{ccccccccc} 0 & \dots & 0 & 1 & \dot{0} & -1 & 0 & \dots & 0 \end{array}\right] \end{array} k^{th}\ \text{row} \quad ; \quad k = 2,\dots,N-1$$

$$B'_1\,\underline{u} = [g_1 e(t), 0, \dots, 0]^T \quad ; \quad B'_k\,\underline{u} = \underline{0} \quad , k = 2,\dots, N$$

The convergence of the proposed synthesis method is best tested by an application to a well known design problem. Let N=5 , d=0 , $g_1 = g_2 = 1$ and consider the problem of finding the elements of $\mathcal{N}(\underline{\varepsilon})$ such that it realizes the impulse response of a fifth order Chebychev filter (1 db ripple). A

Butterworth, low-pass network[15] is used as a starting point. Following the discrete approximation process of Section 3, the specified response[16] is given at 35 equally spaced points ($t_i = ih$; $h = 0.3$ seconds) and an associated P-function (10) is formed with $\eta = 0.1$ and $r_o = 0.0001$. The minimization of this P-function by the Fletcher-Powell method (Section 6) led to the convergence shown on Fig. 3 . Since an exact realization is possible here, the approximation error must reduce to zero. Indeed, after 44 iterations, the minimization of $P(\hat{\underline{\varepsilon}}, r_o)$ led to a point within three significant figures from the point $\hat{\underline{\varepsilon}}^* = [0.4684, 0.9165, 0.33320, 0.9165, 0.4684, 0.0]^T$ which corresponds to a fifth order, 1 db ripple, Chebychev ladder filter.[15] Furthermore, minimization of $P(\hat{\underline{\varepsilon}}; r_\mu)$ for only one value of r_μ was required.

Since the evaluation of $P(\underline{\varepsilon}, \eta, r_\mu)$ and $\nabla P(\underline{\varepsilon}, \eta, r_\mu)$ require the solution of (1) and (2), the greatest part of the necessary computation time is taken by the numerical solution of differential equations. It is, therefore, important that this solution is performed in an efficient manner. When this synthesis approach was programmed (in ALGOL 60) for a UNIVAC 1107 machine, a predictor-corrector method[17] was found very satisfactory. Furthermore, the use of this solution technique extends the range of potential applications of this synthesis method to systems other than linear, time-invariant electrical networks.

REFERENCES

1. Waren, A. D. and L. S. Lasdon, "Practical Filter Design Using Mathematical Optimization," Proc. of the Third Allerton Conference on Circuit and Systems Theory, University of Illinois, Allerton House, Monticello, Illinois, Oct. 20-22, 1965.
2. Bashkow, T. R., "The A-matrix, New Network Description," IRE Trans. Circuit Theory, Vol. CT-4, pp. 117-119, (1957).
3. Bryant, P. R., "The Explicit Form of Bashkow's A-matrix," IRE Trans. Circuit Theory, Vol. CT-9, pp. 303-306, Sept. (1962).
4. Kuh, E. S. and R. A. Rohrer, "The State-Variable Approach to Network Analysis," Proc. IEEE, Vol. 53, July 1965.
5. Rice, J. R., "The Approximation of Functions," Addison-Wesley Publishing Company, Reading, Mass. (1964).
6. Ruston, H., "Synthesis of RLC Networks by Discrete Tschebyscheff Approximations in the Time Domain," Ph.D. Thesis, University of Michigan, 1960.
7. Fischl, R., "Optimal Chebychev Approximation at Discrete Points in the Time Domain," Proc. of the Seventh Midwest Symposium on Circuit Theory, The University of Michigan, Ann Arbor, Michigan, May 4-5, 1964.
8. Tang, D. T., "The Tchebycheff Approximation of a Prescribed Impulse Response with RC Network Realization," IRE Convention Record, Vol. 9, Pt. 4, pp. 214-220, (1961).
9. Lanne, A. A., "Some Approximation Problems in Electrical Circuit Theory," Telecommunications (English translation from Russian) No. 8, Aug. 1965.
10. Zoutendijk, G., "Nonlinear Programming: A Numerical Survey," SIAM Journal on Control, Vol. 4, No. 1, (1966).
11. Fiacco, A. V. and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, A Primal-Dual Method," Management Science, Vol. 10, No. 2, (1964).
12. Lasdon, L. S., Suchman, D. F. and A. D. Waren, "Nonlinear Programming Applied to Linear Array Design," to appear in Acoustical Society Journal, November, 1966.
13. Fletcher, R. and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization,"The Computer Journal (British), Vol. 6, June 1963.
14. Box, M. J., "A Comparison of Several Current Optimization Methods and the Use of Transformations in Constrained Problems," The Computer Journal (British), Vol. 9, May 1966.
15. Weinberg, L., Network Analysis and Synthesis, McGraw-Hill Book Co., New York, (1962).
16. Henderson, K. W. and W. H. Kautz, "Transient Response of Conventional Filters," IRE Trans. on Circuit Theory, December 1958.
17. Hamming, R. N., Numerical Methods for Scientists and Engineers, McGraw-Hill Book Co., New York (1962).

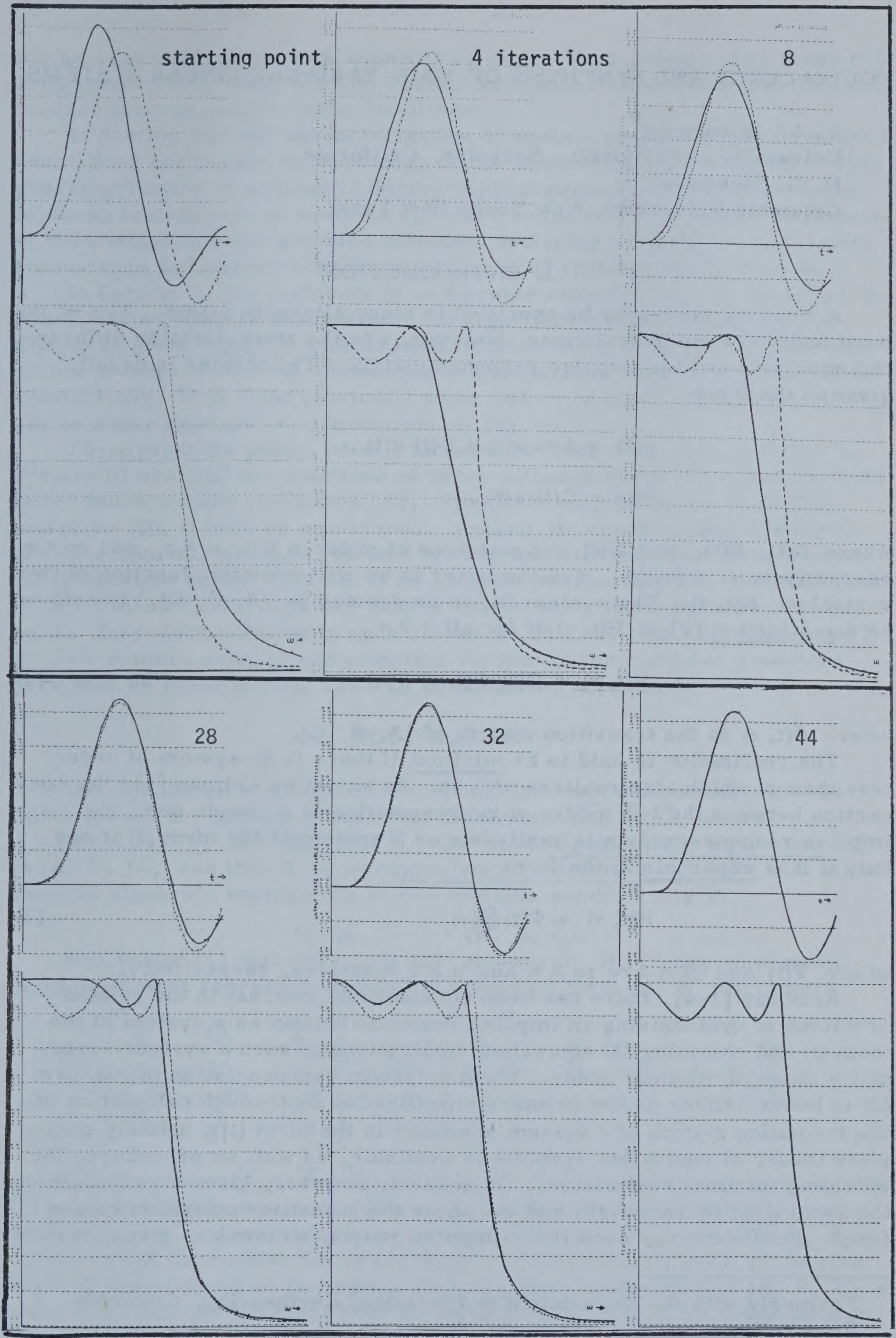


FIGURE 3. The convergence from a Butterworth to a Chebychev filter, impulse response approximation. The associated transmission function (magnitude) is also shown.

EQUIVALENCE AND SYNTHESIS OF TIME-VARIABLE LINEAR SYSTEMS

L. M. Silverman*
University of California, Berkeley, California
H. E. Meadows
Columbia University, New York, New York

1. INTRODUCTION

A linear system may be specified in many alternate forms. Two of the most basic system descriptions, however, are the state-variable differential equations and the impulse response matrix. The former is usually given in the form

$$\begin{aligned} \dot{x}(t) &= A(t)\, x(t) + B(t)\, u(t)\ , \\ y(t) &= C(t)\, x(t)\ , \end{aligned} \tag{1}$$

where A(t), B(t), and C(t) are matrices of order $n \times n$, $n \times r$, and $m \times n$, respectively ($\cdot = (d/dt)$), and the latter is an $m \times r$ matric function of two variables, $H(t, \tau)$. The system (1), to be denoted as (A, B, C),[1] is said to be a realization [1] of $H(t, \tau)$ if for all $t \geq \tau$

$$H(t, \tau) = C(t)\, \Phi(t, \tau)\, B(\tau)\ , \tag{2}$$

where $\Phi(t, \tau)$ is the transition matrix of (A, B, C).

The realization is said to be minimal if there is no system of order less than n which also realizes $H(t, \tau)$. As shown by Kalman [1], the connection between the two modes of representation is a simple one. An impulse response matrix is realizable as a system of the form (1) if and only if it is separable in the form

$$H(t, \tau) = \Psi(t)\, \Theta(\tau) \tag{3}$$

where $\Psi(t)$ and $\Theta(\tau)$ are $m \times n$ and $n \times r$ matrices, respectively.

Recently [1-4], there has been considerable interest in the related problems of synthesizing an impulse response matrix as a system of the form (1) and obtaining all equivalent realizations of such a system--especially those of minimal order. When a system representation in the form (3) is known (either as the primary specification or through calculation of the transition matrix of a system specified in the form (1)), a fairly complete theory of equivalent systems is available, as well as procedures for obtaining minimal realizations. In general, however, given a system in the form (1) it is not possible to calculate the transition matrix in closed form. Furthermore, even if the impulse response matrix is given, it need not be separated or even be known to be separable a priori. For these two important cases, there are few results available in the literature. We will explore them in considerable detail here.

---

* Formerly with the Department of Electrical Engineering, Columbia University.

[1] The explicit dependence of matrices on their argument will be suppressed when the context is clear.

In Section 2 a new characterization of system equivalence is developed which does not depend on knowledge of the impulse response matrix. This characterization is utilized to derive explicit procedures for constructing minimal realizations of systems specified in the form (1). Several results of theoretical interest are also obtained, including criteria for zero-state equivalence and zero-state time-invariance of systems in the form (1).

In Section 3, the synthesis of an impulse response matrix specified in nonseparated form is considered. An algorithm for separating such a matrix is given together with an explicit minimal realization. This realization possesses several advantages in comparison with previous obtained realizations. It is fixed (periodic) when $H(t, \tau)$ is fixed (periodic) and is stable under appropriate constraints on $H(t, \tau)$.

Throughout the paper, it will be assumed that $A(t)$, $B(t)$ and $C(t)$ in system (1) are analytic functions of time, although many of the results hold under much weaker conditions [5]. Similarly only the class of analytic matrices $H(t, \tau)$ will be considered. Analyticity considerably simplifies the development, since, as is well known, analytic matrices have essentially constant rank. That is, if an analytic matrix $G(t)$ has maximal rank k on a finite interval J, it has rank k for all but a finite number of points on J. To reduce notation in such a case we will say that $G(t)$ "has rank k on J." A more general and complete version of the material presented here may be found in [5], and will be published elsewhere.

2. EQUIVALENCE

Let $T(t)$ be an $n \times n$ matrix, nonsingular and continuously differentiable for all t. Then it is said that the system $(\overline{A}, \overline{B}, \overline{C})$ where $\overline{A} = (TA + \dot{T})\,T^{-1}$, $\overline{B} = TB$, and $\overline{C} = CT^{-1}$, is algebraically equivalent [1] to (A, B, C), and that T is an equivalence transformation. We will represent the algebraic equivalence of two systems symbolically as

$$(A, B, C) \xrightarrow{T} (\overline{A}, \overline{B}, \overline{C}).$$

It is known [1] that every analytic system (A, B, C) has an algebraically equivalent representation in the canonical form $(\overline{A}, \overline{B}, \overline{C})$, where

$$\overline{A} = \begin{bmatrix} \overline{A}_{11} & 0 & \overline{A}_{13} \\ \overline{A}_{21} & \overline{A}_{22} & \overline{A}_{23} \\ 0 & 0 & \overline{A}_{33} \end{bmatrix}, \quad \overline{B} = \begin{bmatrix} \overline{B}_1 \\ \overline{B}_2 \\ 0 \end{bmatrix}, \quad \overline{C} = \begin{bmatrix} \overline{C}_1 & 0 & \overline{C}_3 \end{bmatrix} \qquad (4)$$

and the subsystem $(\overline{A}_{11}, \overline{B}_1, \overline{C}_1)$ of order $q \le n$, is completely controllable [1] and completely observable [1]. For analytic systems, complete controllability and observability is equivalent to minimality so that $(\overline{A}_{11}, \overline{B}_1, \overline{C}_1)$ is a minimal realization of (A, B, C). The canonical form (4) is also basic to the problem of generating all zero-state equivalents of (A, B, C) since any such system must have an algebraically equivalent realization in this form, with its minimal subsystem algebraically equivalent to $(\overline{A}_{11}, \overline{B}_1, \overline{C}_1)$. Known procedures [1-3] for obtaining the representation (4) however, rely on the impulse response matrix. To circumvent this and similar difficulties in other system equivalence problems, we propose a new characterization of equivalence which does not require calculation of $H(t, \tau)$.

Consider the sequence of matrices $P_k(t)$ and $S_k(t)$:

$$P_{k+1}(t) = -A(t)\,P_k(t) + \dot{P}_k(t), \quad P_0(t) = B(t) \tag{5}$$

$$S_{k+1}(t) = A'(t)\,\dot{S}_k(t) + S_k(t), \quad S_0(t) = C'(t) \tag{6}$$

For the class of analytic systems it has been shown [5-7] that these matrices effectively characterize controllability and observability. If we define the controllability and observability matrices as

$$Q_c(t) = [P_0(t) \vdots P_1(t) \vdots \cdots \vdots P_{n-1}(t)], \tag{7}$$

$$Q_o(t) = [S_0(t) \vdots S_1(t) \vdots \cdots \vdots S_{n-1}(t)], \tag{8}$$

respectively, then the following criteria hold.

Theorem 1. An analytic system (A, B, C) of order n is completely controllable (observable) if and only if $Q_c(Q_o)$ has rank n.

Because of the intimate connection between controllability, observability and minimality of a system, it is reasonable to expect that the matrices P_k and S_k should also be useful in the minimization problem. That this is indeed the case will now be demonstrated. Let

$$\Gamma_{ij}(t) = \begin{bmatrix} S_0'(t) \\ S_1'(t) \\ \cdot \\ \cdot \\ \cdot \\ S_{i-1}'(t) \end{bmatrix} \begin{bmatrix} P_0(t) \vdots P_1(t) \vdots \cdots \vdots P_{j-1}(t) \end{bmatrix} \tag{9}$$

For reasons made evident by the following theorem, $\Upsilon(t) = \Gamma_{n+1,\,n+1}(t)$ will be termed the equivalence matrix of (A, B, C).

Theorem 2. Two minimum order analytic systems (A, B, C) and $(\overline{A}, \overline{B}, \overline{C})$ are algebraically equivalent if and only if

$$\Upsilon = \overline{\Upsilon}. \tag{10}$$

Proof. Necessity is established by observing that if $(A, B, C) \xrightarrow{T} (\overline{A}, \overline{B}, \overline{C})$ then [8] $\overline{P}_k = TP_k$ and $\overline{S}_k' = S_k T^{-1}$.

To simplify the proof of sufficiency, it will be assumed that $\overline{A} = A = 0$ (this results in no loss of generality since any linear system has an algebraically equivalent representation in this form). If (10) holds, then certainly

$$Q_o' Q_c = \overline{Q}_o' \overline{Q}_c \, . \tag{11}$$

Also, since $P_{k+1} = \dot{P}_k$ if $A = 0$, (10) implies

$$Q_o' \dot{Q}_c = \overline{Q}_o' \dot{\overline{Q}}_c \, . \tag{12}$$

Let n be the order of the two systems, then minimality implies that an interval J exists over which Q_o, Q_c, $\overline{Q}_o$ and $\overline{Q}_c$ have rank n for all t. From (11), therefore, $\overline{Q}_c = TQ_c$, where $T = (\overline{Q}_o \overline{Q}_o')^{-1} \overline{Q}_o \overline{Q}_o'$, is an equivalence transformation on J. It is also clear that $\dot{\overline{Q}}_c = T\dot{Q}_c$. It is immediately deducible from the latter two relationships and the rank of Q_c, that $\dot{T} = 0$ on J. Since the matrices defining T are analytic, it may be concluded that T is constant for all t. Furthermore, it is obvious from (11) and the fact that $\overline{Q}_c = TQ_c$, that $\overline{B} = TB$ and $\overline{C}' = CT^{-1}$, which establishes the theorem.

In [8], methods for reducing the order of a linear system were derived. The specialization of these results to analytic systems leads to

Theorem 3. (a) An analytic system (A, B, C) of order n is algebraically equivalent to a system in the reduced form $(\tilde{A}, \tilde{B}, \tilde{C})$, where

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} \tilde{B}_1 \\ 0 \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} \tag{13}$$

and the q_c-order ($q_c \leq n$) subsystem $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1)$ is completely controllable, if and only if Q_c has rank q_c.

(b) If Q_c has rank q_c, an equivalence transformation which takes (A, B, C) to the form (13) may be found as the solution to

$$TQ_c = \begin{bmatrix} \tilde{Q}_{c1} \\ 0 \end{bmatrix} \tag{14}$$

where $\tilde{Q}_{c1}$ is an arbitrary analytic matrix having q_c rows and rank equal to that of Q_c.

Several explicit transformations based on (14) have been given [8] for reducing a system to the form (13).

Following from Theorem 3 (together with its dual form in terms of Q_o) is an explicit measure of the degree of the minimum order realizations of an analytic system and a constructive procedure for generating its minimal realizations.

Theorem 4. An analytic system (A, B, C) is algebraically equivalent to a system of the form (4) (and thus zero-state equivalent to a minimal system of order q) if and only if Γ_{nn} has rank q.

Proof. Necessity is obvious from the form of system (4).

Sufficiency is established by exhibiting a transformation which takes (A, B, C) to the form (4).[2] An outline of this procedure follows. Let $q_c \leq n$ be the rank of Q_c. From Theorem 3, a transformation T_c may be found which reduces (A, B, C) to the form (13). Let q_o be the rank of the observability matrix of $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1)$. Using the dual of Theorem 3, we construct T_o such that $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1)$ is transformed to a system having coefficient matrices of the form

$$\begin{bmatrix} \overline{A}_{11} & O \\ \overline{A}_{21} & \overline{A}_{22} \end{bmatrix}, \quad \begin{bmatrix} \overline{B}_1 \\ \overline{B}_2 \end{bmatrix}, \quad \begin{bmatrix} \overline{C}_1 & O \end{bmatrix}, \tag{15}$$

and the subsystem $(\overline{A}_{11}, \overline{B}_1, \overline{C}_1)$, of order q_o, is completely observable (it obviously must be completely controllable). Thus the combined transformation

$$T = \begin{bmatrix} T_o & O \\ O & I \end{bmatrix} T_c$$

takes (A, B, C) to the canonical form (4). If $\overline{Q}_{cl}$ and $\overline{Q}_{ol}$ are the controllability and observability matrices of $(\overline{A}_{11}, \overline{B}_1, \overline{C}_1)$, then it is clear that $\Gamma_{q_o q_o} = \overline{Q}'_{ol}\,\overline{Q}_{cl}$. Thus $\Gamma_{q_o q_o}$, and consequently Γ_{nn}, has rank q_o. Therefore, $q_o = q$, which completes the proof.

Let (A, B, C) be any minimal system of order q with equivalence matrix Υ, and let $(\tilde{A}, \tilde{B}, \tilde{C})$ be any system of order $n \geq q$, zero-state equivalent to (A, B, C). Then it is obvious from the above proof that

$$\Upsilon = \tilde{\Gamma}_{q+1,\,q+1} \tag{16}$$

This relationship leads to a strong generalization of Theorem 2.

Corollary 1. Let (A, B, C) and $(\overline{A}, \overline{B}, \overline{C})$ be analytic systems of order n and $\overline{n}$, respectively. Then the systems are zero-state equivalent if and only if Γ_{nn} and $\overline{\Gamma}_{\overline{n}\overline{n}}$ have the same rank q and $\Upsilon = \overline{\Upsilon}$, where $\Upsilon = \Gamma_{q+1,\,q+1}$ and $\overline{\Upsilon} = \overline{\Gamma}_{q+1,\,q+1}$.

The equivalence matrix thus characterizes all realizations of an impulse response matrix. Furthermore, as will now be shown, relation (16) implies that a minimal order system can be found directly from the equivalence matrix of a nonminimal realization.

Let Υ be the equivalence matrix of a class of minimal systems of order q, and isolate the submatrices: (a) $\Gamma = \Gamma_{qq}$, the matrix formed from the first mq rows and the first rq columns of Υ; (b) Γ^*, the matrix formed from the last mq rows and the first rq columns of Υ; (c) F, any $q \times q$ submatrix of Γ having rank q for all t; (d) F^*, that submatrix of Γ^* occupying the same position in Γ^* as F in F^*; (e) F_1, the matrix composed of those columns of the first m rows of Γ which correspond to the columns of F; (f) F_2, the matrix composed of those rows of the first r columns of Γ which correspond to the rows of F.

---

[2] Independently, a similar transformation for the special case of fixed systems was recently given by Anderson et al. [9].

Observe that the matrices F, F^*, F_1 and F_2 are all submatrices of Υ and can be determined without an a priori knowledge of any minimal realization.

Theorem 5. Let (A, B, C) be an analytic system of order n and let $\Upsilon = \Gamma_{q+1,\, q+1}$ be the equivalence matrix of the minimal order realizations of (A, B, C). Then for any matrix F satisfying (c) above, the system $(F^* F^{-1}, F_2, F_1 F^{-1})$, is a minimal realization of (A, B, C), where F^*, F_1 and F_2 are defined by (d), (e) and (f), respectively.

The proof of this theorem will be deferred to the following section. Let us remark, however, that this method for reducing a system to a least order zero-state equivalent, is computationally superior to the transformational procedure of Theorem 4 [5]. Furthermore, it leads to several results of considerable theoretical interest. For example,

Corollary 2. Let (A, B, C) be an analytic system of order n. Then (A, B, C) is zero-state time-invariant (i.e., has a stationary impulse response matrix) if and only if Γ_{nn} has rank $q \leq n$ and $\Upsilon = \Gamma_{q+1,\, q+1}$ is a constant matrix.

To prove the corollary, merely note that if Υ is constant, then the system $(F^* F^{-1}, F_2, F_1 F^{-1})$ as defined in Theorem 5 is a constant system.

3. SYNTHESIS

Let $H(t, \tau)$ be an $m \times r$ matric function analytic in t and τ, and define

$$\Gamma_{ij}(t, \tau) = \begin{bmatrix} H_{00}(t, \tau) & H_{01}(t, \tau) & \cdots & H_{0,\, j-1}(t, \tau) \\ H_{10}(t, \tau) & H_{11}(t, \tau) & \cdots & H_{1,\, j-1}(t, \tau) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ H_{i-1,\, 0}(t, \tau) & H_{i-1,\, 1}(t, \tau) & \cdots & H_{i-1,\, j-1}(t, \tau) \end{bmatrix}, \tag{17}$$

where $H_{ij}(t, \tau) = (\partial^i/\partial t^i)(\partial^j/\partial \tau^j)\, H(t, \tau)$. If $H(t, \tau)$ is realized by system (A, B, C), it may be readily verified that

$$H_{ij}(t, \tau) = S_i'(t)\, \Phi(t, \tau)\, P_j(\tau)\, . \tag{18}$$

Thus, $\Gamma_{ij}(t, t) = \Gamma_{ij}(t)$ as defined by (9) for the system (A, B, C). This relationship will be utilized to derive a method for synthesizing $H(t, \tau)$ which does not require an a priori assumption of separability in the form (3). It is based in part on a method of Borskii [10] for realizing a scalar impulse response function as an input-output differential equation.

Theorem 6. $H(t, \tau)$ is separable in the form (3) if and only if there exists a finite positive integer n such that $\Gamma_{nn}(t, \tau)$ and $\Gamma_{n+1, n+1}(t, \tau)$ both have rank n for $t \geq \tau$. Furthermore, n is the order of the minimal realizations of $H(t, \tau)$.

Proof. Let $\Gamma^*_{nn}(t, \tau) = (\partial/\partial t)\, \Gamma_{nn}(t, \tau)$, then $\Gamma^*_{nn}(t, \tau)$ is the matrix formed from the last mn rows of $\Gamma_{n+1, n+1}(t, \tau)$. Also, let $F(t, \tau)$ be any $n \times n$ submatrix of $\Gamma_{nn}(t, \tau)$ having rank n and let $F^*(t, \tau)$ be the corresponding submatrix of $\Gamma^*_{nn}(t, \tau)$. If we define

$$A(t, \tau) = F^*(t, \tau)\, F^{-1}(t, \tau)\,, \tag{19}$$

then

$$\Gamma^*_{nn}(t, \tau) = A(t, \tau)\, \Gamma_{nn}(t, \tau) \tag{20}$$

wherever $F(t, \tau)$ has maximal rank. The most important property of $A(t, \tau)$ as defined by (19), is that it is not a function of τ. To see this, differentiate (19) with respect to τ, and observe that (20) implies

$$\frac{\partial}{\partial \tau} F^*(t, \tau) = A(t, \tau) \frac{\partial}{\partial \tau} F(t, \tau)\,. \tag{21}$$

We now show that $F(t, \tau)$ is separable in the form

$$F(t, \tau) = F(t, \lambda)\, F^{-1}(\xi, \lambda)\, F(\xi, \tau)\,, \tag{22}$$

where (ξ, λ) is any point at which $F(\xi, \lambda)$ is nonsingular. With minor manipulation, the relationship

$$\frac{\partial}{\partial t}\left[F^{-1}(t, \lambda)\, F(t, \tau) \right] = -F^{-1}(t, \lambda) \left[A(t, \lambda) - A(t, \tau) \right] F(t, \tau) \tag{23}$$

may be obtained. Since $A(t, \tau) = A(t, \lambda)$ it is clear that $F^{-1}(t, \lambda)\, F(t, \tau)$ is not a function of t. Thus, for any (ξ, λ) such that $F(\xi, \lambda)$ is nonsingular, $F^{-1}(t, \lambda)\, F(t, \tau) = F^{-1}(\xi, \lambda)\, F(\xi, \tau)$, and (20) follows. It is a simple matter to show that separability of $F(t, \tau)$ implies separability of $H(t, \tau)$. Since $H(t, \tau)$ is a submatrix of $\Gamma_{nn}(t, \tau)$ there must exist matrices $C_1(t, \tau)$ and $B_1(t, \tau)$ such that

$$H(t, \tau) = C_1(t, \tau)\, F(t, \tau)\, B_1(t, \tau)\,. \tag{24}$$

Let $F_1(t, \tau)$ $(F_2(t, \tau))$ be the matrix composed of those columns (rows) of the first m rows (r columns) of $\Gamma_{nn}(t, \tau)$ which correspond to the columns (rows) of $F(t, \tau)$. It may be verified that $C_1(t, \tau) = F_1(t, \tau)\, F^{-1}(t, \tau)$ and

$B_1(t, \tau) = F^{-1}(t, \tau)\, F_2(t, \tau)$. Precisely as for $A(t, \tau)$, it may be shown that $C_1(t, \tau)$ is not a function of τ, and similarly that $B_1(t, \tau)$ is not a function of t. Using this information it is easily established that

$$H(t, \tau) = F_1(t, \lambda)\, F^{-1}(\xi, \lambda)\, F_2(\xi, \tau)\,. \tag{25}$$

If we let $\Psi(t) = F_1(t, \lambda)$ and $\Theta(t) = F^{-1}(\xi, \lambda)\, F_2(\xi, t)$ then clearly, (O, Θ, Ψ) is an nth order realization of $H(t, \tau)$. Since $\Gamma_{nn}(t, t) = \Gamma_{nn}(t)$ has rank n, it is also apparent that (O, Θ, Ψ) is a minimal realization. This completes the proof of sufficiency. Necessity is obvious: if (A, B, C) is an nth-order minimal realization $\Gamma_{nn}(t, \tau)$ and $\Gamma_{n+1, n+1}(t, \tau)$, must have rank n since

$$\Gamma_{nn}(t, \tau) = Q'_o(t)\, \Phi(t, \tau)\, Q_c(\tau)\,. \tag{26}$$

Although the realization procedure in the above proof does not require an initially separated impulse response matrix, it still shares the major drawback of the standard procedure [1-3]--it results in a realization having a zero feedback matrix. This type of realization, because of its inherent instability, is quite undesirable practically. Furthermore, for fixed or periodic impulse response matrices, the corresponding realizations will not in general be fixed or periodic. As will now be shown, however, the separation algorithm can be modified so as to yield physically acceptable realizations.

If $F(t, \tau)$ has rank n for all $t \geq \tau$ then (with a slight abuse of notation)

$$A(t) = A(t, \tau) = F^{*}(t, t)\, F^{-1}(t, t) = F^{*}(t)\, F^{-1}(t)$$

$$C_1(t) = C_1(t, \tau) = F_1(t, t)\, F^{-1}(t, t) = F_1(t)\, F^{-1}(t)$$

$$B_1(\tau) = B_1(t, \tau) = F^{-1}(\tau, \tau)\, F_2(\tau, \tau) = F^{-1}(\tau)\, F_2(\tau)$$

are continuous matric functions. Consequently, we may express $H(t, \tau)$ as

$$H(t, \tau) = C_1(t)\, \Phi(t, \tau)\, F(\tau, \tau)\, B_1(\tau) \tag{27}$$

where $\Phi(t, \tau) = F(t, \lambda)\, F^{-1}(\tau, \lambda)$ is the transition matrix of $A(t)$. Thus $H(t, \tau)$ is realized by the system $(F^{*} F^{-1}, F_2, F_1 F^{-1})$. Observe that this realization is identical to that given in Theorem 5, so that the above derivation also constitutes a proof of that theorem. Although not unique (save in the single-output case) this realization is quite "canonical" in nature. It may be readily verified that if $H(t, \tau) = H(t - \tau)$ the realization will be constant and if $H(t, \tau) = H(t + s, \tau + s)$ for some s the realization will be periodic with period s.

In general, under appropriate conditions on $H(t, \tau)$, it may also be established that the realization given above will be stable (in the sense of Lyapunov). This important aspect of the realization problem will be discussed more fully in a forthcoming paper [11].

ACKNOWLEDGMENT

The research reported herein was supported in part by the National Science Foundation under Grant NSF GP-2789, in part by the Office of Naval Research under Contract Nonr-4259(04) (both to Columbia University, New York) and in part by the Joint Services Electronics Program (U. S. Army, U. S. Navy and U. S. Air Force) under Grant AF-AFOSR-139-66 and in part by the National Science Foundation under Grant GK-716 (both grants to the Electronics Research Laboratory, University of California, Berkeley).

REFERENCES

[1] R. E. Kalman, "Mathematical description of linear dynamical systems," J. SIAM Control, Vol. 1, pp. 152-192; 1963.

[2] R. E. Kalman and L. Weiss, "Contributions to linear system theory," Int. J. Eng. Sci., Vol. 3, pp. 141-171, Pergamon Press, 1965.

[3] D. C. Youla, "The synthesis of linear dynamical systems from prescribed weighting patterns," J. SIAM, Vol. 14, pp. 527-549, May 1966.

[4] C. A. Desoer and P. Varaiya, "The minimal realization of a nonanticipative impulse response matrix," University of California, Berkeley, Tech Memo. M-138, Dec. 16, 1965.

[5] L. M. Silverman, "Representation and realization of time-variable linear systems," Technical Report No. 94, Department of Electrical Engineering, Columbia University, June 1966.

[6] A. R. Stubberud, "A controllability criterion for a class of linear systems," IEEE Trans. on Appl. and Ind., Vol. 68, pp. 411-413, November 1964.

[7] A. Chang, "An algebraic characterization of controllability," IEEE Trans. on Automatic Control, Vol. AC-10, pp. 112-113, Jan. 1965.

[8] L. M. Silverman and H. E. Meadows, "Degrees of controllability in time-variable linear systems," Proc. of N.E.C., Vol. 21, pp. 689-693, 1965

[9] B. D. O. Anderson, R. W. Newcomb, R. E. Kalman, and D. C. Youla, "Equivalence of linear time-invariant systems," J. Franklin Inst., Vol. 281, pp. 371-378, May 1965.

[10] V. Borskii, "On the properties of impulsive response functions with variable parameters," Automation and Remote Control, Vol. 20, pp. 822-830, July 1959.

[11] L. M. Silverman, "Stable realization of impulse response matrices," to be presented at the IEEE Intl. Convention, March 1967.

CONTROLLABILITY AND OBSERVABILITY OF COMPOSITE SYSTEMS

C. T. CHEN and C. A. DESOER
University of California, Berkeley, California

ABSTRACT

In this paper a complete characterization of the controllability and observability of some natural representations of composite systems is presented. These systems are obtained by connecting in parallel, in tandem, and in a feedback loop, linear time-invariant finite dimensional systems. Emphasis is placed on the distinction between a system and one of its representations. Some problems treated by Gilbert [1] for the case of distinct eigenvalues are completely solved here. In addition we establish the equivalence, in a very broad setting, of the controllability and the observability of a feedback connection to that of a tandem connection: it is interesting to note that for controllability this equivalence applies to nonlinear systems. In order to remain within the page number limitation, the proofs are omitted here but will be published later.

SYSTEMS AND THEIR REPRESENTATIONS

The basic notion associated with a system is the set of all its input-output pairs: this is the set of all inputs, $\underset{\sim}{u}$, and outputs $\underset{\sim}{y}$ where $\underset{\sim}{u}$ and $\underset{\sim}{y}$ are functions of time defined on $[t_0, \infty)$ where the initial time t_0 may vary from pair to pair and ranges over the whole real line. Given the set of all pairs $(\underset{\sim}{u}, \underset{\sim}{y})$, we assume that it is possible to assign a set of parameters called the state. It is assumed that this parameterization satisfies all the consistency requirements [2] and is such that the output $\underset{\sim}{y}$ is a function of the input $\underset{\sim}{u}$ and the initial state. Given a system, there are many ways of assigning a state to it and, consequently, it is important to keep in mind the difference between a system and its representation. Therefore, throughout this paper we make a sharp distinction between a system and one of its representations. A representation of a system is a set of rules that allows the calculation of the output $\underset{\sim}{y}$ on the basis of the initial state and of the input $\underset{\sim}{u}$.

This work was supported by the Joint Services Electronics Program (U. S. Army, U. S. Navy, and U. S. Air Force) under Grant AF-AFOSR-139-65, and by the National Science Foundation under Grant GK-716.

In this paper we consider almost exclusively differential systems that are linear and time invariant. Typically, we consider a system $\mathcal{S}$ that has the following representation S :

$$S : \dot{\underset{\sim}{x}} = \underset{\sim}{A}\,\underset{\sim}{x} + \underset{\sim}{B}\,\underset{\sim}{u}\,, \tag{2.1a}$$

$$\underset{\sim}{y} = \underset{\sim}{C}\,\underset{\sim}{x} + \underset{\sim}{D}\,\underset{\sim}{u}\,, \tag{2.1b}$$

where $\underset{\sim}{x}$, the state, is an n-dimensional vector; the input $\underset{\sim}{u}$, a $p \times 1$ vector; the output $\underset{\sim}{y}$, a $q \times 1$ vector. $\underset{\sim}{A}$, $\underset{\sim}{B}$, $\underset{\sim}{C}$, and $\underset{\sim}{D}$ are, respectively, $(n \times n)$, $(n \times p)$, $(q \times n)$, and $(q \times p)$ matrices. The zero-state response of $\mathcal{S}$ is completely characterized by the $q \times n$ matrix (called the transfer function matrix)

$$\underset{\sim}{\hat{G}}(s) \triangleq \underset{\sim}{C}(s\underset{\sim}{I} - \underset{\sim}{A})^{-1}\,\underset{\sim}{B} + \underset{\sim}{D}\,. \tag{2.2}$$

Indeed

$$\underset{\sim}{\hat{y}}(s) = \underset{\sim}{\hat{G}}(s)\,\underset{\sim}{\hat{u}}(s)\,, \tag{2.3}$$

where $\underset{\sim}{\hat{u}}(s)$ and $\underset{\sim}{\hat{y}}(s)$ are the Laplace transforms of the input $\underset{\sim}{u}$ and of the corresponding zero-state response $\underset{\sim}{y}$. From its definition, $\underset{\sim}{\hat{G}}(s)$ is independent of the chosen parametrization of the input-output pairs. It should be stressed that $\underset{\sim}{\hat{G}}(s)$ does not characterize $\mathcal{S}$, indeed it characterizes only the zero-state responses of $\mathcal{S}$ [2]. In terms of the Kalman decomposition [3] $\underset{\sim}{\hat{G}}(s)$ characterizes only the part that is both controllable and observable.

In order to test the representation of composite systems, we start by considering differential systems with representations S_1 and S_2, respectively:

$$S_i : {}_i\dot{\underset{\sim}{x}} = {}_i\underset{\sim}{A}\,{}_i\underset{\sim}{x} + {}_i\underset{\sim}{B}\,{}_i\underset{\sim}{u} \tag{2.4a}$$

$${}_i\underset{\sim}{y} = {}_i\underset{\sim}{C}\,{}_i\underset{\sim}{x} + {}_i\underset{\sim}{D}\,{}_i\underset{\sim}{u} \tag{2.4b}$$

where ${}_i\underset{\sim}{A}$, ${}_i\underset{\sim}{B}$, ${}_i\underset{\sim}{C}$, and ${}_i\underset{\sim}{D}$ are ${}_in \times {}_in$, ${}_in \times {}_ip$, ${}_iq \times {}_in$, and ${}_iq \times {}_ip$ matrices, respectively; ${}_i\underset{\sim}{u}$ is a ${}_ip \times 1$ input vector, ${}_i\underset{\sim}{y}$ is a ${}_iq \times 1$ output vector, and ${}_i\underset{\sim}{x}$ is an ${}_in \times 1$ state vector. A few words about notations: in general, the subscript on the left-hand side denotes the system (e.g.,

${}_i\underset{\sim}{x}$, ${}_i\underset{\sim}{A}$, ${}_ip$). Subscripts on the right-hand side of these letters will be required in the next section. It is well known that Eq. (2.4a) has a unique solution for ${}_i\underset{\sim}{x}$ and that ${}_i\underset{\sim}{x}$ qualifies as the state of $\mathcal{S}_i$. Let Σ_i be the state space of the representation S_i, thus ${}_i\underset{\sim}{x} \in \Sigma_i$. In the following when we connect systems $\mathcal{S}_1$ and $\mathcal{S}_2$ -- either in parallel, tandem, or feedback -- we use in the representation of the composite system $\Sigma_1 \oplus \Sigma_2$ as state space, equivalently, $[{}_1\underset{\sim}{x}, {}_2\underset{\sim}{x}]'$ as state.

One approach to the study of controllability is the following. Consider the tandem connection of $\mathcal{S}_1$ followed by $\mathcal{S}_2$. Clearly $[{}_1\underset{\sim}{x}, {}_2\underset{\sim}{x}]'$ qualifies as the state of the tandem representation. And

$$\begin{bmatrix} {}_1\hat{\underset{\sim}{x}}(s) \\ \\ {}_2\hat{\underset{\sim}{x}}(s) \end{bmatrix} = \begin{bmatrix} (s\,{}_1\underset{\sim}{I} - {}_1\underset{\sim}{A})^{-1}\,{}_1\underset{\sim}{B} \\ \\ (s\,{}_2\underset{\sim}{I} - {}_2\underset{\sim}{A})^{-1}\,{}_2\underset{\sim}{B}[{}_1\underset{\sim}{C}(s\,{}_1\underset{\sim}{I} - {}_1\underset{\sim}{A})^{-1}\,{}_1\underset{\sim}{B} + {}_1\underset{\sim}{D}] \end{bmatrix} {}_1\hat{\underset{\sim}{u}}(s) \qquad (3.1)$$

It is well known that the state $[{}_1\underset{\sim}{x}, {}_2\underset{\sim}{x}]'$ is controllable on any subinterval of $[0,\infty)$ if and only if the rows of the matrix in Eq. (3.1) are linearly independent (more precisely, if and only if there is no nonzero $({}_1n + {}_2n)$-tuple of real numbers $(\xi_1, \xi_2, \cdots, \xi_{{}_1n+{}_2n}) = \underset{\sim}{\xi}$ such that the product of $\underset{\sim}{\xi}$ by the matrix of Eq. (3.1) is not an identically zero row-vector). This characterization of the controllability of the state $[{}_1\underset{\sim}{x}, {}_2\underset{\sim}{x}]'$ does not give any insight into the following phenomenon: examples show that it is possible that the representations S_1 and S_2 are controllable and observable, that the state ${}_1\underset{\sim}{x}$ of S_1 and the state ${}_2\underset{\sim}{x}$ of S_2 are controllable (separately) by the input ${}_1\underset{\sim}{u}$ of the tandem connection, but that the state $[{}_1\underset{\sim}{x}, {}_2\underset{\sim}{x}]'$ of the tandem connection is not controllable.

In order to gain some insight into these phenomena, we shall use the Jordan canonical form for the representation of each system. The properties of controllability and observability being invariant under nonsingular transformations of the state, there is no loss of generality in assuming that $\underset{\sim}{A}$ is in the Jordan canonical form. Let $\underset{\sim}{A}$ be of the form shown in Table 1: $\underset{\sim}{A}$ is an $n \times n$ matrix in Jordan canonical form with $m(m \leq n)$ distinct eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_m$. Let A_i denote all the Jordan blocks associated with the eigenvalue λ_i and $r(i)$ be the number of Jordan blocks in $\underset{\sim}{A}_i$. Let $\underset{\sim}{A}_{ij}$ be the $\underline{j}$ th Jordan block in $\underset{\sim}{A}_i$, then $\underset{\sim}{A}_i = \text{diag.}(\underset{\sim}{A}_{i1}, \underset{\sim}{A}_{i2}, \cdots, \underset{\sim}{A}_{ir(i)})$ and

$\underset{\sim}{A} = \text{diag.}(\underset{\sim}{A}_1, \underset{\sim}{A}_2, \cdots, \underset{\sim}{A}_m)$. Let n_i and n_{ij} be the order of $\underset{\sim}{A}_i$ and $\underset{\sim}{A}_{ij}$, respectively; then

$$n = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} \sum_{j=1}^{r(i)} n_{ij} .$$

$\underset{\sim}{B}$ and $\underset{\sim}{C}$ are arbitrary $n \times p$, $q \times n$ matrices. Corresponding to $\underset{\sim}{A}_i$ and $\underset{\sim}{A}_{ij}$, $\underset{\sim}{B}$ and $\underset{\sim}{C}$ are partitioned as shown in Table 1. Call $\underset{\sim}{b}_{1ij}$, $\underset{\sim}{b}_{\ell ij}$ the first and the last row of $\underset{\sim}{B}_{ij}$; $\underset{\sim}{c}_{1ij}$, $\underset{\sim}{c}_{\ell ij}$ the first and the last column of $\underset{\sim}{C}_{ij}$.

Theorem 3.1 (stated without proof by Kalman [5]). The representation S is controllable if and only if for each $i = 1, 2, \cdots, m$, the set of $r(i)$ p-dimensional row vectors

$$\underset{\sim}{b}_{\ell i1}, \underset{\sim}{b}_{\ell i2}, \cdots, \underset{\sim}{b}_{\ell ir(i)}$$

is a linearly independent set.

Applying the duality theory of Kalman, we have

Theorem 3.2. The representation S is observable if and only if, for each $i = 1, 2, \cdots, m$, the set of $r(i)$ q-dimensional column vectors

$$\underset{\sim}{c}_{1i1}, \underset{\sim}{c}_{1i2}, \cdots, \underset{\sim}{c}_{1ir(i)}$$

is a linearly independent set.

PARALLEL CONNECTION

Define $\Lambda_i \triangleq \{{}_i\lambda_j \mid j = 1, 2, \cdots, m_i\}$ for $i = 1, 2$.

Theorem 4.1. Assume that the representations S_1 and S_2 are controllable and observable, and that ${}_1p = {}_2p$, ${}_1q = {}_2q$. Then the parallel connection of S_1 and S_2 is controllable (observable) if and only if either $\Lambda_1 \cap \Lambda_2 = \phi$ or, in case Λ_1 and Λ_2 are not disjoint, each pair of common eigenvalues, say ${}_1\lambda_\alpha = {}_2\lambda_\beta$, is such that the set of $(r_1(\alpha) + r_2(\beta))$ ${}_1p$ - dimensional row vectors ${}_1\underset{\sim}{b}_{\ell\alpha 1}, {}_1\underset{\sim}{b}_{\ell\alpha 2}, \cdots, {}_1\underset{\sim}{b}_{\ell\alpha r_1(\alpha)}, {}_2\underset{\sim}{b}_{\ell\beta 1}, {}_2\underset{\sim}{b}_{\ell\beta 2}, \cdots, {}_2\underset{\sim}{b}_{\ell\beta r_2(\beta)}$ is a linearly independent set. (The set of $(r_1(\alpha) + r_2(\beta))$ ${}_1q$ - dimensional column vectors ${}_1\underset{\sim}{c}_{1\alpha 1}, {}_1\underset{\sim}{c}_{1\alpha 2}, \cdots, {}_1\underset{\sim}{c}_{1\alpha r_1(\alpha)}, {}_2\underset{\sim}{c}_{1\alpha r_1(\alpha)}, {}_2\underset{\sim}{c}_{1\beta 1}, {}_2\underset{\sim}{c}_{1\beta 2}, \cdots, {}_2\underset{\sim}{c}_{1\beta r_2(\beta)}$ is a linearly independent set.)

Table 1

Notations for the Jordan Form Representation

$$\underset{\sim}{A} = \begin{bmatrix} \underset{\sim}{A}_1 & & & & \\ & \underset{\sim}{A}_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \underset{\sim}{A}_m \end{bmatrix} \qquad \underset{\sim}{B} = \begin{bmatrix} \underset{\sim}{B}_1 \\ \underset{\sim}{B}_2 \\ \cdot \\ \cdot \\ \underset{\sim}{B}_m \end{bmatrix}$$

$$(n \times n) \qquad\qquad (n \times p)$$

$$\underset{\sim}{C} = [\underset{\sim}{C}_1 \ \underset{\sim}{C}_2 \ \cdots \ \underset{\sim}{C}_m]$$

$$\underset{\sim}{A}_i = \begin{bmatrix} \underset{\sim}{A}_{i1} & & & & \\ & \underset{\sim}{A}_{i2} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \underset{\sim}{A}_{ir(i)} \end{bmatrix} \qquad \underset{\sim}{B}_i = \begin{bmatrix} \underset{\sim}{B}_{i1} \\ \underset{\sim}{B}_{i2} \\ \cdot \\ \cdot \\ \underset{\sim}{B}_{ir(i)} \end{bmatrix}$$

$$(n_i \times n_i) \qquad\qquad (n_i \times p)$$

$$\underset{\sim}{C}_i = [\underset{\sim}{C}_{i1} \ \underset{\sim}{C}_{i2} \ \cdots \ \underset{\sim}{C}_{ir(i)}]$$

$$(q \times n_i)$$

$$\underset{\sim}{A}_{ij} = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \cdot & \cdot & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{bmatrix} \qquad \underset{\sim}{B}_{ij} = \begin{bmatrix} \underset{\sim}{b}_{1ij} \\ \underset{\sim}{b}_{2ij} \\ \cdot \\ \cdot \\ \underset{\sim}{b}_{\ell ij} \end{bmatrix}$$

$$(n_{ij} \times n_{ij})$$

$$\underset{\sim}{C}_{ij} = [\underset{\sim}{c}_{1ij} \ \underset{\sim}{c}_{2ij} \ \cdots \ \underset{\sim}{c}_{\ell ij}]$$

The proof is a strightforward application of Theorems 3.1 and 3.2.

Corollary 4.1. Assume that S_1 and S_2 are controllable and observable and that ${}_1p = {}_1q = {}_2p = {}_2q = 1$. Then the parallel connection of S_1 and S_2 is controllable and observable if and only if $\Lambda_1 \cap \Lambda_2 = \phi$.

TANDEM CONNECTION

Case I. No common eigenvalues: $\Lambda_1 \cap \Lambda_2 = \phi$.

We give first the characterization.

Theorem 5.1. Assume that the representations S_1 and S_2 are controllable and observable, and that ${}_1q = {}_2p$. If $\Lambda_1 \cap \Lambda_2 = \phi$, then S_{12}, the state representation of the tandem connection of S_1 followed by S_2, is controllable (observable) if and only if, for each $i = 1, 2, \cdots, m_2$, the set of $r_2(i)\,{}_1p$ - dimensional row-vectors

$$ {}_2\underset{\sim}{b}_{\ell i1}\,\hat{\underset{\sim}{G}}_1({}_2\lambda_i),\ {}_2\underset{\sim}{b}_{\ell i2}\,\hat{\underset{\sim}{G}}_1({}_2\lambda_i),\ \cdots,\ {}_2\underset{\sim}{b}_{\ell ir_2(i)}\,\hat{\underset{\sim}{G}}_1({}_2\lambda_i) $$

is a linearly independent set. (For each $i = 1, 2, \cdots, m_1$, the set of $r_1(i)\,{}_2q$ - dimensional column-vectors

$$ \hat{\underset{\sim}{G}}_2({}_1\lambda_i)\,{}_1\underset{\sim}{c}_{1i1},\ \hat{\underset{\sim}{G}}_2({}_1\lambda_i)\,{}_1\underset{\sim}{c}_{1i2},\ \cdots,\ \hat{\underset{\sim}{G}}_2({}_1\lambda_i)\,{}_1\underset{\sim}{c}_{1ir_1(i)} $$

is a linearly independent set.)

The proof uses in a key step the identity

$$ \frac{1}{s - {}_2\lambda_i}(s\underset{\sim}{I} - {}_1\underset{\sim}{A})^{-1} = \frac{1}{s - {}_2\lambda_i}({}_2\lambda_i\underset{\sim}{I} - {}_1\underset{\sim}{A})^{-1} - ({}_2\lambda_i\underset{\sim}{I} - {}_1\underset{\sim}{A})^{-1}(s\underset{\sim}{I} - {}_1\underset{\sim}{A})^{-1} $$

Sufficient condition can be expressed in a form that is independent of the representation:

Theorem 5.2. Assume that S_1 and S_2 are controllable and observable and that ${}_1q = {}_2p$. If $\Lambda_1 \cap \Lambda_2 = \phi$ and if

$\hat{\underset{\sim}{G}}_1({}_2\lambda_i)$ has rank ${}_1q$ for all ${}_2\lambda_i \in \Lambda_2$ and

$(\hat{\underset{\sim}{G}}_2({}_1\lambda_i)$ has rank ${}_2p$ for all ${}_1\lambda_i \in \Lambda_1)$,

then S_{12} is controllable (observable).

Case II. $\Lambda_1 \cap \Lambda_2 \neq \phi$.

We give here only one result in order to give an idea of the nature of the additional conditions required.

Theorem 5.3. Assume that S_1 and S_2 are controllable and observable, and that ${}_1q = {}_2p$. If Λ_1 and Λ_2 are not disjoint, then for each pair of common eigenvalues, say ${}_1\lambda_\alpha = {}_2\lambda_\beta$, we assume that (i) $r_1(\alpha) = 1$, $r_2(\beta) = 1$; i.e., corresponding to ${}_1\lambda_\alpha = {}_2\lambda_\beta$, there is only one Jordan block in ${}_1\underset{\sim}{A}$ and ${}_2\underset{\sim}{A}$; and (ii) ${}_2\underset{\sim}{b}_{\ell\beta 1}\,{}_1\underset{\sim}{c}_{1\alpha 1} \neq 0$. Under these assumptions, S_{12} is controllable (observable) if and only if, for each $i = 1, 2, \cdots, \beta - 1, \beta + 1, \cdots, m_2$, the set of $r_2(i)\,{}_1p$ - dimensional row-vectors

$$ {}_2\underset{\sim}{b}_{\ell i1}\,\hat{\underset{\sim}{G}}_1({}_2\lambda_i),\ {}_2\underset{\sim}{b}_{\ell i2}\,\hat{\underset{\sim}{G}}_1({}_2\lambda_i),\ \cdots,\ {}_2\underset{\sim}{b}_{\ell i r_2(i)}\,\hat{\underset{\sim}{G}}_1({}_2\lambda_i) $$

FEEDBACK CONNECTIONS

We propose to show that for a very general class of systems the problem of controllability and observability of a feedback connection reduces to that of the tandem connection. As an example of the class of system to which these results apply we consider nonlinear time-varying systems with finite-dimensional differential equation representations:

$$ \overline{S}_i : {}_i\dot{\underset{\sim}{x}} = {}_i\underset{\sim}{f}({}_i\underset{\sim}{x}, {}_i\underset{\sim}{u}, t), \quad {}_i\underset{\sim}{y} = {}_i\underset{\sim}{g}({}_i\underset{\sim}{x}, {}_i\underset{\sim}{u}, t) \tag{6.1} $$

where ${}_i\underset{\sim}{x}$, ${}_i\underset{\sim}{u}$ and ${}_i\underset{\sim}{y}$ are, respectively, ${}_in \times 1$, ${}_ip \times 1$ and ${}_iq \times 1$ vectors; ${}_i\underset{\sim}{f}$ is an ${}_in$ vector-valued function, ${}_i\underset{\sim}{g}$ is a ${}_iq$ vector-valued function. The state space of $\overline{S}_i$ is denoted by Σ_i and the input space is assumed to be a linear vector space. (We write $\overline{S}_i$ instead of S_i to indicate that we are considering nonlinear systems.) $\overline{S}_i$ is said to be determinate [2, p. 96] if for any ${}_i\underset{\sim}{u}$, any t_0 and any initial state ${}_i\underset{\sim}{x}(t_0)$, there exist unique ${}_i\underset{\sim}{x}(t)$ and $\underset{\sim}{y}_i(t)$ for all $t \geq t_0$. It is well known [10] that $\overline{S}_i$ is determinate if ${}_i\underset{\sim}{f}$ is Lipschitz in ${}_i\underset{\sim}{x}$ and continuous in $\underset{\sim}{u}_i$ and t and ${}_i\underset{\sim}{g}$ is continuous in ${}_i\underset{\sim}{x}$, ${}_i\underset{\sim}{u}$, and t.

For a determinate representation $\overline{S}_i$, since ${}_i\underset{\sim}{x}(t)$ and ${}_i\underset{\sim}{y}(t)$ are functions of the input and the initial state, all the concepts of controllability and observability can be applied with appropriate modifications. As in the linear case, the nonlinear representation $\overline{S}_i$ is said to be controllable on

$[t_0, t_1]$ if for any given state ${}_i\underset{\sim}{x}(t_0) \in \Sigma_i$, there exists an input $\underset{\sim}{u}_i$ which transfers ${}_i\underset{\sim}{x}(t_0)$ to the zero state at time $t_1 < \infty$. $\overline{S}_i$ is said to be *observable on $[t_0, t_1]$* if for any given input $\underset{\sim}{u}$ over $[t_0, t_1]$, the response $\underset{\sim}{y}$ (due to this input $\underset{\sim}{u}$ and an arbitrary initial state $\underset{\sim}{x}(t_0)$) and the knowledge of the representation (6.1) suffice to determine that initial state $\underset{\sim}{x}(t_0)$. Similarly, $\overline{S}_i$ is said to be *zero-input observable on $[t_0, t_1]$* if, with $\underset{\sim}{u}(t) \equiv 0$, the response $\underset{\sim}{y}$ over $[t_0, t_1]$ (due to an arbitrary initial state $\underset{\sim}{x}(t_0)$) and the knowledge of the representation (6.1) suffice to determine the initial state $\underset{\sim}{x}(t_0)$. No explicit criterion for the controllability and observability of such determinate representation $\overline{S}_i$ is known. However it turns out that if a feedback connection is determinate, the representation of the feedback connection is controllable (observable) if and only if some appropriate open-loop connection is controllable (observable).

Theorem 6.1. We are given two determinate systems with representations $\overline{S}_1$ and $\overline{S}_2$ of the form (6.1) with ${}_1p = {}_2q$, ${}_1q = {}_2p$. Let $\overline{S}_f$, the representation of the feedback connection, (with $\overline{S}_1$ in the forward path, $\overline{S}_2$ in the feedback path, and negative feedback), be determinate and $\Sigma_1 \oplus \Sigma_2$ be the state space of $\overline{S}_f$. Under these assumptions, $\overline{S}_f$ is controllable on $[t_0, t_1]$ if and only if $\overline{S}_{12}$ is controllable on $[t_0, t_1]$ ($\overline{S}_{12}$ is $\overline{S}_1$ followed by $\overline{S}_2$). Furthermore, $\overline{S}_f$ is zero-input observable on $[t_0, t_1]$ if $\overline{S}'_{21}$ is observable on $[t_0, t_1]$, where $\overline{S}'_{21}$ is the tandem connection of $\overline{S}_2$, a sign inverter, and $\overline{S}_1$.

REFERENCES

[1] E. G. Gilbert, "Controllability and observability in multivariable control systems," *SIAM J. Control*, vol. 1, pp. 128-151 (1963).

[2] L. A. Zadeh and C. A. Desoer, *Linear System Theory*, McGraw-Hill Book Co., Inc., New York (1963).

[3] R. E. Kalman, "Canonical structure of linear dynamical systems," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 48, No. 4, pp. 496-600 (1962).

[4] G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, Revised Ed., The MacMillan Company, New York (1953).

[5] R. E. Kalman, "Irreducible realizations and the degree of a matrix of rational functions," *SIAM J.*, vol. 13, pp. 520-544 (1965).

[6] E. Kreindler and P. E. Sarachik, "On the concepts of controllability and observability of linear systems," *IEEE Trans. on Automatic Control*, vol. AC-9, pp. 129-136 (1964).

[7] R. E. Kalman, "On the general theory of control systems," Proc. First Intl. Cong. Automatic Control, Moscow, 1960, vol. 1, pp. 481-493, Butterworth and Co., Ltd., London (1961).

[8] P. L. Falb and M. Athans, "A direct constructive proof of the criterion for complete controllability," IEEE Trans. on Automatic Control, vol. AC-9, pp. 180-190 (1964).

[9] F. R. Gantmacher, The Theory of Matrices, vol. 1, Chelsea, New York (1959).

[10] G. Birkhoff and G. Rota, Ordinary Differential Equations, Ginn and Company, Boston (1962).

ON RELATIONSHIPS BETWEEN THE THEORIES OF DISCRETE-STATE AND CONTINUOUS-STATE LINEAR TIME-INVARIANT SYSTEMS*

P. H. SWAIN and K. S. FU
Purdue University
Lafayette, Indiana

ABSTRACT

Recent investigations in the fields of continuous-state systems (eg., differential systems) and discrete-state systems (eg., sequential machines) have made increasingly apparent the following observation: When the restrictions of time-invariance and linearity are imposed, the analyses and results for both discrete-state and continuous-state systems become strikingly similar. But while several authors have noted parallel results in these areas, little effort has been made to exploit this parallelism. Thus, the purpose of this paper is twofold: (1) to summarize for future reference some existing results which are applicable to both continuous-state and discrete-state systems, and (2) to use these results to make advances in the particular areas where they have not yet been applied.

The use of state equations for system description is discussed briefly, leading to equivalence of systems, canonical forms, transforms and transfer functions. The dual concepts of controllability and observability are presented in forms suitable for application to both discrete-state and continuous-state systems. The property of strong-connectedness of sequential machines is explicitly related to controllability.

1. INTRODUCTION

A few years ago, R. E. Kalman [1] noted the similarity between a theorem he had formulated and a theorem due to E. F. Moore [2]. Kalman's theorem dealt with controllability and observability of linear differential systems; Moore's was concerned with strongly-connected automata and indistinguishable states. "Evidently", wrote Kalman, "the two theorems are concerned with the same abstract facts, each being stated in a different mathematical framework".

Interest has been growing in such parallels between the theories of continuous-state and discrete-state systems and in the possibility of developing a common unified theory [3,4,5]. Recently pioneering efforts have been made in this direction, most notably by M. Arbib [4,5]. In this paper we exploit in some detail the common aspects of the theories of linear, time-invariant, differential systems and linear, time-invariant sequential systems. The restriction to linear, time-invariant systems is made only because the most interesting results to date have been obtained for this class of systems. It is obvious, however, that some of these results are immediately extendable to a more general class of systems.

2. NOTATION AND DEFINITIONS

Several authors [1,5,6] have given definitions for systems (and related concepts) which are general enough to encompass all of the systems of interest here. We give a brief summary of essential definitions, with only as much detail as our development will require.

2.1 System

A system is a mathematical structure $S = (U,Y,X,\lambda,\delta)$ consisting of
a state space X,
an input space U consisting of admissible inputs to the system, and an
input segment space U^*. An input segment $u(t_o,t]$ is a mapping
$u: (t_o,t] \to U^*$,

---

*This work was supported by the National Science Foundation Grant GK-696.

an output space Y,
a set of values of time, T, at which the behavior of the system is defined,
a state transition function λ: $X \times U^* \times T \to X$,
an output function δ: $X \times U \times T \to Y$.

That is, for any t_o, $t \in T$ ($t_o \leq t$), any initial state $x(t_o) \underline{\Delta}\, x_o \in T$, and any admissible input segment $u(t_o,t]$, we have

$$\begin{aligned} x(t) &= \lambda(x_o,\ u(t_o,t],\ t_o) \\ y(t) &= \delta(x(t),\ u(t),\ t) \end{aligned} \tag{2.1}$$

Furthermore, for any $t_1 \in T$ satisfying $t_o \leq t_1 \leq t$, and any fixed input segment $u(t_o,t]$, the following hold:

$$\begin{aligned} &\lambda(x_o,\ u(t_o,t_o],\ t_o) = x_o \\ &\lambda(x_o,\ u(t_o,\ t],\ t_o) = \\ &\qquad \lambda(\lambda(x_o,\ u(t_o,t_1],t_o),\ u(t_1,t],\ t_1) \end{aligned} \tag{2.2}$$

Note that under this definition a system is inherently nonanticipatory; i.e., any input following time τ does not affect the behavior of the system before τ.

It is possible, at least in principle, to eliminate $x(t)$ in equations (2.1). The result has the form

$$y(t) = F(x_o,\ u(t_o,t],\ t) \tag{2.3}$$

This will be called the input-output equation of the system, whereas (2.1) will be referred to as the state equations of the system.

2.2 Time-invariance

A system is time-invariant if $u(t_o,t_1] = u(t_o + \tau,\ t_1 + \tau]$ implies

$$F(x_o,u(t_o,t_1],\ t_1) = F(x_o,\ u(t_o + \tau,\ t_1 + \tau],\ t_1 + \tau)$$

for all initial states $x_o \in X$, all $u \in U^*$, all t_o, $t_1 \in T$, and all $\tau \geq 0$ such that $t_o + \tau$, $t_1 + \tau \in T$. In this case, the functions F, λ, δ do not depend explicitly on time. Since the systems of interest in this paper are time-invariant, mention of the time variable will be deleted from the arguments of the functions.

2.3 Linearity

A state z of system S is a zero state if

$$y(t) = F(z,0_t) = 0$$

for all $t \in T$, where 0_t is the zero input segment of duration t.

In the next definition, assume all inputs in each equation (including the zero input) have equal duration.

A system S is linear if it has the following three properties:

The decomposition property

$$F(x_o,u) = F(x_o,0) + F(z,u) \tag{2.4a}$$

for all $x_o \in X$, $u \in U^*$,

Zero-state linearity

$$F(z,k(u-v)) = kF(z,u) - kF(z,v) \tag{2.4b}$$

for all u, $v \in U^*$, and all k such that the products indicated are defined,

Zero-input linearity

$$F(k(x_1-x_2),0) = kF(x_1,0) - kF(x_2,0) \tag{2.4c}$$

for all x_1, $x_2 \in X$, and all k such that the products indicated are defined.

2.4 Differential and sequential systems

For definitness, we will deal with two common types of systems. Linear time-invariant differential systems, the behavior of which is described by the vector differential equations

$$\begin{aligned} \dot{x} &= Ax + Bu, \quad x(t_o) = x_o \\ y &= Cx + Du \end{aligned} \tag{2.5}$$

are familiar to control theorists. In equation (2.5), $x=x(t)$ is an n-dimensional state vector ($\dot{x} = dx/dt$), $y = y(t)$ is a p-dimensional output vector, $u=u(t)$ is an r-dimensional input vector, and A, B, C, D are nxn, nxr, pxn, and pxr constant matrices, respectively. Linear sequential machines, described by the vector difference equations

$$\begin{aligned} x' &= Ax + Bu, \quad x(t_o) = x_o \\ y &= Cx + Du \end{aligned} \tag{2.6}$$

are familiar to automata and coding theorists. Except that $x'=x'(t)=x(t+1)$, the symbols in (2.6) have the same meaning as those in (2.5).

The important differences between the two types of systems are: (1) For differential systems, T is taken as an interval of the real line, usually $T = [o, \infty)$; whereas the behavior of sequential systems is considered on a quantized time scale, say $T = (0, 1, 2, 3, \ldots)$. (2) The input, output, and state spaces of differential systems are Euclidean spaces (whence their classification as continuous-state); whereas the input, output and state spaces of sequential systems consist of discrete (usually finite) sets (whence their classification as discrete-state). The important similarities of the systems described by (2.5) and (2.6) (at least as far as this paper is concerned) are linearity and time-invariance.

3. COMMON THEORY FOR LINEAR, TIME-INVARIANT SYSTEMS

3.1 State equations

First we show that linear, time-invariant differential and sequential systems can be described within the framework of the definition of a system given in Section 2. These results are well known [5]. For the differential system, the state equations are (taking $t_o=0$)

$$\begin{aligned} x(t) &= \lambda(x_o, u(0,t]) \\ &= \Phi(t)\, x_o + \int_0^t \Phi(t-\tau)\, B\, u(\tau) d\tau \\ y(t) &= \delta(x(t), u(t)) \\ &= C\, x(t) + D\, u(t) \end{aligned} \tag{3.1}$$

where the transition matrix Φ satisfies

$$\dot{\Phi}(t) = A\, \Phi(t), \qquad \Phi(o) = I \tag{3.2}$$

For the sequential system (recall $T = (0, 1, 2, 3, \ldots)$)

$$\begin{aligned} x(t) &= \lambda(x_o, u(0,t]) \\ &= A^t\, x_o + \sum_{m=0}^{t-1} A^{t-m-1} B\, u(m) \\ y(t) &= \delta(x(t), u(t)) \\ &= C\, x(t) + D\, u(t) \end{aligned} \tag{3.3}$$

Note that the form of equations (3.1) and (3.3) is the same.

Since equations (2.5) and (2.6) are equivalent to equations (3.1) and (3.3) in the sense that either set may be derived from the other, both sets will be referred to as the state equations of the corresponding system. Where we wish to state a result which applies equally well to either system (2.5) or (2.6), we will state the result once and use $\overline{x}(t)$ to stand for $x(t)$ and $x'(t) = x(t+1)$. Thus,

$$\begin{aligned} \overline{x}(t) &= A\, x(t) + B\, u(t) \\ y(t) &= C\, x(t) + D\, u(t) \end{aligned} \tag{3.4}$$

are the state equations of a linear, time-invariant (differential or sequential) system.

3.2 Equivalent states and systems; canonical forms.

Two states x_R and x_S, belonging to systems R and S respectively, are equivalent states, $x_R \equiv x_S$, if

$$F(x_R,u) = F(x_S,u) \tag{3.5}$$

for all $u \in U^*$ (R and S may or may not refer to the same system, but in any case the systems must have the same input set U and output set Y).

A system is in reduced form if there are no distinct states in its state space which are equivalent.

R and S are equivalent systems, $R \equiv S$, if for every state of R there is an equivalent state of S and vice versa.

In the remainder of the paper, it will be assumed that the input, output and state sets are vector spaces over a field $\mathcal{F}$. This represents no significant practical restriction with respect to differential systems. An important class of sequential systems which obey this restriction is linear binary machines (also n-ary machines if n is a prime integer).

Let T be a nonsingular matrix such that

$$\hat{x} = Tx \tag{3.6}$$

This implies a one-to-one relationship between the state spaces $\hat{\mathbf{X}}$ and $\mathbf{X}$, i.e., a simple change of coordinates. Substituting (3.6) into (3.4) we get

$$\begin{aligned} \hat{x} &= TAT^{-1}\,\hat{x} + TB\,u = \hat{A}\hat{x} + \hat{B}u \\ y &= CT^{-1}\,\hat{x} + Du = \hat{C}\hat{x} + \hat{D}u \end{aligned} \tag{3.7}$$

The new system will, for any given input, have the same output as the original system if they are both started in initial states related by $\hat{x}_o = T\,x_o$. Thus, the systems are equivalent, as are any other systems related to these according to

$$\begin{aligned} \hat{A} &= TAT^{-1} & \hat{C} &= CT^{-1} \\ \hat{B} &= TB & \hat{D} &= D \end{aligned} \tag{3.8}$$

for some nonsingular matrix T. A corrolary of this result is: Given any system of the form (3.4) and a matrix $\hat{A}$ which is similar to A (i.e., $A = TAT^{-1}$ for some nonsingular T), there exists a system given by (3.8) which is equivalent to the original system.

This last result allows us to study the properties of any system of the form (3.4) by examining the properties of the matrix A, hereafter referred to as the state matrix. It also suggests the possibility of finding a system which is in some sense the simplest equivalent of a given system. These ideas are closely related, as indicated in the development to follow.

If A is an n x n matrix and if the degree of its minimum polynomial m(x) is n, then A is said to be nonderogatory (otherwise derogatory).

Theorem [7,11]: Canonical forms. Every nonderogatory matrix A is similar to the companion matrix of its miminum polynomial; every derogatory matrix is similar to a matrix of the form

$$C = \begin{bmatrix} C_1 & 0 \cdot \cdot \cdot \cdot \cdot & 0 \\ 0 & C_2 \cdot \cdot \cdot \cdot \cdot & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & 0 \cdot \cdot \cdot \cdot \cdot \cdot & C_m \end{bmatrix} \tag{3.9}$$

where the C_i are the companion matrices of the elementary divisors of A.

The essential properties of a linear system are invariant under the (similarity) transformation which takes the state matrix into the canonical forms defined by the theorem; therefore, much can be learned about an apparently complex system by studying the simpler canonical form. Furthermore, structural relationships hidden in the given system may be revealed by the coordinate change in the state space; eg., a state matrix of the form C in (3.9) exhibits decoupling between various sets of states. Of course, how much simpler (if at all) the transformed system will be depends on the original system and on the field over which the state vector space is defined.

3.3 Transforms; transfer functions.

Transforms and transfer functions are important tools in the analysis and design of linear differential systems. Work by Huffman [8,9] and Gill [10] has demonstrated the usefulness of these concepts in dealing with linear binary systems. We proceed to generalize their ideas slightly and fill in some details, using as a guide the results obtained for linear differential systems. To avoid cumbersome notational problems, it will be convenient to work with single input, single output systems. There is no problem extending the results to multiple input/output systems.

Let the input and output sequences of a linear sequential machine consist of elements from a field $\mathcal{F}$. Any such sequence, say u: $u_o, u_1, u_2, \ldots$, where u_i is the input at time t_i, can be expressed as a polynomial in D with coefficients over $\mathcal{F}$.

$$u(D) = u_o + Du_1 + D^2u_2 + \ldots + D^iu_i + \ldots \tag{3.10}$$

D is considered to be a delay operator or delayor. Multiplication by D will delay the sequence by one interval of time; multiplication by D^{-1} will advance the sequence by one interval. The formal power series (3.10) will be called the delay transform of the sequence u.

It has been shown [8,9,11] that the behavior of a sequential machine whose output is a function of the present and a finite number of previous inputs and a finite number of previous outputs can be described by the ratio of two polynomials in the delay operator D. This ratio

$$T(D) = h(D)/g(D) \tag{3.11}$$

is called the transfer function (or transfer ratio) of the linear sequential machine. If the machine is in the zero state initially, then its response sequence y to an input sequence u is given by the coefficients of the polynomial

$$y(D) = T(D)\, u(D) \tag{3.12}$$

Many useful properties of transforms and transfer functions which are known for continuous systems apply equally well in the discrete case. For example, in synthesizing a system, if only the zero-state response is required, then minimization can be carried out by dividing the greatest common factor out of the numerator and denominator of the transfer function.

The relationship between the state equations and the transfer function of a linear time-invariant differential system is well known (see [1], for example). The aim of the remainder of this section is to find a general expression which gives the transfer function of any linear system of the form

$$\begin{aligned} \overline{x}(t) &= Ax(t) + b\, u(t) \\ y(t) &= c^Tx(t) + d\, u(t) \end{aligned} \tag{3.13}$$

where the symbols are interpreted as before, except that b, c, and d are vectors. Superscript T denotes the transpose of a vector. We assume that the relation between the transforms of x and $\overline{x}$ can be expressed in the form

$$g\,\overline{x} = x \qquad \text{or} \qquad \overline{x} = g^{-1}x \tag{3.14}$$

where g is any linear operator or transform variable.

An interesting way to achieve the desired result is by means of signal flow analysis. We make use of the following:

1. Every linear system is equivalent to a system of which the state matrix has the canonical form described in the theorem in Section 3.2. Thus it is only necessary to consider companion matrices as system matrices. (We will later show that the transfer functions of equivalent systems are identical).

2. A linear signal flow graph is a graphical means of representing linear equations.

3. The gain of a linear flow graph is given by Mason's formula [6]:

$$G = \frac{\sum_i G_i \Delta_i}{\Delta} \tag{3.15}$$

where the sum is taken over all forward paths from input to output; G_i is the gain of the ith forward path; Δ is the graph determinant given by

$$\Delta = 1 - \sum_r P_{r1} + \sum_s P_{s2} - \sum_t P_{t3} + \ldots \tag{3.16}$$

where P_{r1} is the loop gain of the rth loop, with the summation over all loops; P_{s2} is the product of the loop gains of the sth pair of nontouching loops, the summation taken over all such pairs; P_{t3} is the product of the loop gains of the tth set of three nontouching loops, the summation taken over all such sets; etc. Δ_i is the determinant of the graph obtained from the original graph by deleting all branches of the ith forward path and all branches having a node in common with it.

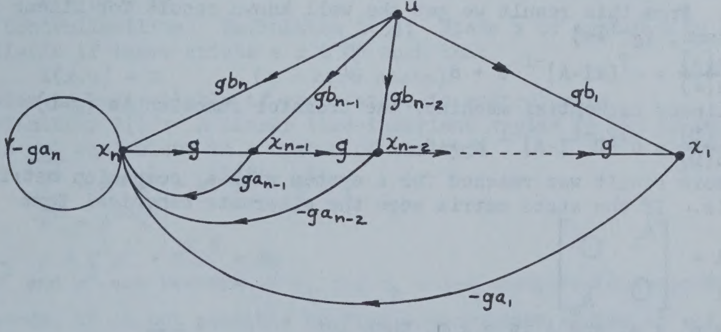
Let the system in question be given by the state equations

$$\bar{x} = \frac{1}{g} x = \begin{bmatrix} 0 & 1 & 0 \ldots\ldots & 0 & 0 \\ 0 & 0 & 1 \ldots\ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 \ldots\ldots & 0 & 1 \\ -a_1 & -a_2 & -a_3 & -a_{n-1} & -a_n \end{bmatrix} x + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} u \tag{3.17}$$

$$y = [C_1 \ C_2 \ \ldots \ C_{n-1} \ C_n]x + du \tag{3.18}$$

It will suffice to solve (3.17) for x in terms of u and substitute the result in equation (3.18) (it is understood that we are dealing with transformed variables).

One possible signal flow graph for the set of n equations (3.17) is shown below.



First we consider x_1 as the output variable. The expression for the graph determinant is simplified by the fact that there are no nontouching loops in the graph. We have

$$\Delta = 1 - \sum_r P_{r1}$$

$$= 1 + ga_n + g^2 a_{n-1} + \ldots + g^n a_1$$

Now if we let $M = g^{-1}I - A$ = characteristic matrix of A (with indeterminate g^{-1}), laborious but straightforward computation reveals that

$$\Delta = g^n|g^{-1}I - A| = g^n \det M$$

Furthermore,

$$G_i\Delta_i = (g^i b_i)(1+ga_n + \ldots + g^{n-i}a_{i+1})$$

$$= b_i g^n \operatorname{cof} M_{i1}$$

where cof M_{i1} is the cofactor of the $(i,1)$ element of M. Summing over i yields

$$\frac{x_1}{u} = \frac{g^n \sum_{i=1}^{n} b_i \operatorname{cof} M_{i1}}{g^n \det M} = \frac{\sum_{i=1}^{n} b_i \operatorname{cof} M_{i1}}{\det M}$$

Following the same procedure for the remaining x_i, $i=2,3,\ldots n$, the result is

$$\frac{x}{u} = \frac{1}{\det M} \cdot \begin{bmatrix} \sum_{i=1}^{n} b_i \operatorname{cof} M_{i1} \\ \vdots \\ \sum_{i=1}^{n} b_i \operatorname{cof} M_{in} \end{bmatrix}$$

$$= M_{CT}\, b/\det M$$

where M_{CT} is the cofactor matrix of the transpose of M. Since $M_{CT}/\det M = M^{-1}$, we have

$$x = M^{-1}\, bu = [g^{-1}\, I - A]^{-1}\, bu$$

and

$$y = c^T[g^{-1}I - A]^{-1}\, bu + du$$

i.e.,

$$\frac{y}{u} = c^T[g^{-1}I - A]^{-1}\, b + d$$

is the transfer function of the linear time-invariant system given by (3.17) and (3.18). From this result we get the well known result for linear differential systems, $(g^{-1}=s)$

$$\frac{y(s)}{u(s)} = c^T[sI - A]^{-1}\, b + d$$

and for a linear sequential machine, the transfer function is $(g=D)$

$$\frac{y(D)}{u(D)} = c^T[D^{-1}I - A]^{-1}\, b + d$$

The above result was reached for a system with a **companion** matrix for state matrix. If the state matrix were the alternate canonical form

$$A = \begin{bmatrix} A_1 & & O \\ & \ddots & \\ O & & A_m \end{bmatrix}$$

where A_i is $n_i \times n_i$ and A is $n \times n$, then let

$$b = \begin{bmatrix} b_1 \\ \cdots \\ \vdots \\ \cdots \\ b_m \end{bmatrix} \begin{matrix} \} \; n_1 \text{ elements} \\ \\ \\ \\ \} \; n_m \text{ elements} \end{matrix} \qquad c = \begin{bmatrix} c_1 \\ \cdots \\ \vdots \\ \cdots \\ c_m \end{bmatrix} \begin{matrix} \} \; n_1 \\ \\ \\ \\ \} \; n_m \end{matrix}$$

It can be shown that

$$\frac{y}{u} = \sum_{i=1}^{m} \left\{ c_i^T\, [g^{-1}I_{n_i} - A_i]^{-1}\, b_i \right\} + d$$

where I_{n_i} is the $n_i \times n_i$ identity matrix.

Let a system be related to one of the canonical forms by $\hat{x} = Tx$

$$\hat{A} = TAT^{-1} \qquad \hat{c}^T = c^T T^{-1}$$
$$\hat{b} = Tb \qquad \hat{d} = d$$

Then we find

$$\begin{aligned} y/u &= c^T[g^{-1}I-A]^{-1}b + d \\ &= \hat{c}^T\, T[g^{-1}I-T^{-1}\hat{A}T]^{-1}\, T^{-1}\hat{b} + \hat{d} \\ &= \hat{c}^T\, [g^{-1} I-\hat{A}]^{-1}\, \hat{b} + \hat{d} \end{aligned}$$

Thus, transfer functions are invariant under similarity transformations on the state matrix; i.e. equivalent systems have identical transfer functions.

Other instances where the theory of linear binary machines has been extended to parallel the theory of differential systems may be found in Gill [10].

3.4 Controllability and Observability

Kalman [12,1] showed that the concepts of controllability and observability are vital to the study of control theory. Besides being essential to the description and prediction of internal and output behavior of systems, these concepts lend considerable insight as to the basic structural properties of systems. It is significant that the mathematical machinery developed through investigation of controllability and observability has proven applicable to the study of many difficult problems, such as the control of multivariable systems [13].

Because they give various insights to the implications of controllability and observability, we present some different definitions for these concepts, asserting that, for the types of systems discussed in this paper, those for controllability are equivalent and those for observability are equivalent. It should be clear that within the framework of our "common linear, time-invariant systems theory" these concepts are equally applicable to continuous-state and discrete-state systems.

Again we refer to systems of the form

$$\begin{aligned} \bar{x} &= Ax + Bu, \quad x(t_o) = x_o \\ y &= Cx + Du \end{aligned} \tag{3.19}$$

3.4.1 Controllability. Definition 1[5]: State x of system S is said to be controllable if there exists a $u \in U^*$ such that

$$\lambda(x,u) = z \qquad (z = \text{zero state}) \tag{3.20}$$

The system is controllable if every state is controllable.

Definition 2[1]: A linear time-invariant system (3.19) is controllable if it is not equivalent to a system of the type

$$\begin{aligned} \bar{x}^1 &= A^{11}x^1 + A^{12}x^2 + B^1u \\ \bar{x}^2 &= A^{22}x^2 \\ y &= C^1x^1 + C^2x^2 + Du \end{aligned} \tag{3.21}$$

where x^1 and x^2 are vectors of n_1 and $n_2 = n-n_1$ components respectively. (In other words, it is not possible to find a coordinate system in which the state variables are separated into two groups such that the second group is not affected by the first group or by the system inputs).

We then have the following theorem which may be proved for systems of the form (3.19) by using Definition 1 above and equations (3.1) and (3.3).

Theorem [6]: The system S given by (3.19) is controllable if and only if the matrix

$$[B, AB, \ldots, A^{n-1}B]$$

has rank n (i.e. the column vectors of this matrix span the state space X).

We are almost ready to give the explicit relationship between controllability and strong-connectedness mentioned in the introduction. First we have two definitions.

A state x_2 of S is *reachable* from a state x_1 of S if there exists an input segment $u \in U^*$ such that $x_2 = \lambda(x_1,u)$.

A system is *strongly-connected* if every state is reachable from every other state.

Now we can assert: a linear time-invariant system S is strongly-connected if and only if it is controllable. To see this observe that for linear systems strong-connectedness always implies controllability. On the other hand, in a controllable system the zero state is reachable from every other state, and in a linear system every state is reachable from the zero state [6, Chapter 11]. Therefore, every state is reachable from every other state.

An excellent treatment of some special aspects of the controllability of linear sequential machines may be found in [14].

3.4.2 Observability. Definition 1 [5]: A linear time-invariant system S is observable if and only if it is in reduced form.

Definition 2 [1]: a linear, time-invariant system (3.19) is observable if it is *not* equivalent to a system of the form

$$\begin{aligned} \bar{x}^1 &= A^{11}x^1 + B^1u \\ \bar{x}^2 &= A^{21}x^1 + A^{22}x^2 + B^2u \qquad (3.22) \\ y &= C^1x^1 + Du \end{aligned}$$

(symbols as in Section 3.4.1) (In other words, it is not possible to find a coordinate system in which the state variables are separated into two groups, such that the second group does not affect either the first group or the outputs of the system).

Definition 3 [6]: A system S given by (3.19) is said to be observable if for some $T > 0$ and all initial states $x_o \in X$, the knowledge of A, C, and the zero input response $y[0,T]$ suffices to determine the initial state x_o.

Using Definition 3 and equations (3.1) and (3.3) with $u \equiv 0$, we can establish the following theorem.

Theorem [6]: The system S given by (3.19) is observable if and only if the matrix

$$[C^*, A^*C^*, \ldots, A^{*(n-1)}C^*]$$

has rank n (* denotes the complex conjugate of the transpose of the matrix).

As implied by the above definitions, the concepts of observable system, reduced form, and initial state determinability are essentially the same.

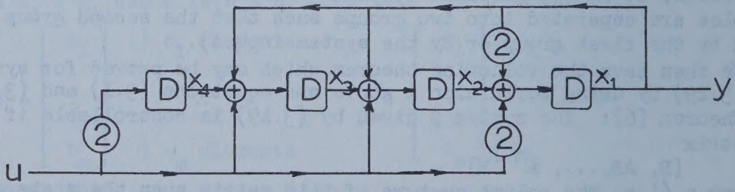
4. EXAMPLE

Consider the following linear sequential system with ternary logic (all arithmetic operations performed modulo 3):

$$x' = Ax + bu = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 2 \\ 1 \\ 1 \\ 2 \end{bmatrix} u$$

$$y = c^T x = [1\ 0\ 0\ 0]\ x \qquad (d=0)$$

This system can be synthesized using delayors, mod 3 adders, and constant multipliers:



By applying the results of Section 3.3, we find the transfer function of the system is

$$\frac{Y(D)}{U(D)} = \frac{2D^4 + D^3 + D^2 + 2D}{2D^3 + 2D^2 + D + 1}$$

However, checking the controllability and observability matrices (see theorems in Sections 3.4.1 and 3.4.2) **reveals** that the system is neither controllable nor observable. This implies that the numerator and denominator of the transfer function have common factors [1]. Assuming that we are in fact dealing with a synthesis (rather than analysis) problem, the unobservable parts of the system are useless, which is also the case, with a few exceptions (such as a clock circuit), for the uncontrollable parts. We can eliminate these parts by cancelling the common numerator and denominator factors. To do this, we can avoid the problem of actually factoring these expressions by using the Euclidean algorithm to determine the greatest common divisor of numerator and denominator. This divisor is removed by ordinary long division of polynomials. In the present case, the greatest common divisor is found to be D^2+2 (remember: mod 3 arithmetic) and the simplified transfer function is

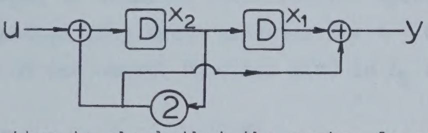
$$T(D) = \frac{D^2 + 2D}{D+1} = \frac{D(D+2)}{D+1}$$

Clearly, no common factors remain. One possible set of state equations for the reduced system is

$$x' = \begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

$$y = [1\ 2]\, x$$

which may be synthesized in the form



It is a trivial matter to check that the system described by the "reduced" state equations is both controllable and observable.

5. CONCLUDING REMARKS

The method of this paper has been chiefly to take well established results for continuous-state linear time-invariant systems and show that they are equally valid in the discrete-state case. An increasing number of investigators are taking advantage of the reverse relationship by applying the techniques of the automata theorist to continuous-state systems. It appears that these opportunities to try fresh approaches where the old ones have been pretty well exhausted will continue to result in significant advances in both of these branches of systems theory.

REFERENCES

1. R.E. Kalman, "Mathematical Description of Linear Dynamical Systems", Journal S.I.A.M., Ser. A, Vol. 1, pp. 152-192; 1963.
2. E.F. Moore, "Gedanken Experiments on Sequential Machines", Automata Studies, Princeton University Press, Princeton, N.J., 1956.
3. B.S. Morgan, "Multivariable Systems", 1965 I.E.E.E. International Convention Record, pt. 6, pp. 87-95.
4. M.A. Arbib, "Automata Theory and Control Theory - A Rapproachement", Automatica, Vol. 3, pp. 161-189, Pergamon Press, 1966.
5. M.A. Arbib, "A Common Framework for Automata Theory and Control Theory", Journal S.I.A.M. on Control, Ser. A, Vol. 3, pp. 207-222; 1965.
6. L.A. Zadeh and C.A. Desoer, Linear System Theory: The State Space Approach, McGraw-Hill Book Co., Inc., New York, N.Y.; 1963.
7. S. Perlis, Theory of Matrices, Addison-Wesley Press, Inc., Cambridge, Mass.; 1952.
8. D.A. Huffman, "The Synthesis of Linear Coding Networks", Information Theory, C. Cherry, ed., Academic Press, New York, N.Y.; 1953.

9. D.A. Huffman, "A Linear Circuit Viewpoint on Errorcorrecting Codes", IRE Trans. on Information Theory, Vol. IT-2, pp. 20-28; September, 1956.
10. A. Gill, Introduction to the Theory of Finite-State Machines, McGraw-Hill Book Co., Inc., New York, N.Y.; 1962.
11. W.W. Peterson, Error Correcting Codes, M.I.T. Press, Cambridge, Mass.; 1961.
12. R.E. Kalman, "On the General Theory of Control Systems", Proc. First International Congress on Automatic Control, Vol. 1, pp. 481-492, Butterworths, London; 1961.
13. R.E. Kalman, "On Structural Properties of Linear, Constant, Multivariable Systems", I.F.A.C., paper 6.A.; June, 1966.
14. M. Cohn, "Controllability in Linear Sequential Networks", IRE Trans. on Circuit Theory, Vol. CT-9, pp. 74-78; March, 1962.

HILBERT TRANSFORMS AND GENERALIZED FUNCTIONS.

HARRY SCHWARZLANDER
Department of Electrical Engineering, Syracuse University
Syracuse, New York.

Abstract

The definition of the Hilbert transform is extended in such a way that it applies within the framework of generalized functions. Hilbert transforms of generalized functions are defined in terms of suitable weakly-convergent sequences in the space of generalized functions. A generalized function $\overline{x(t)}$ which corresponds to an ordinary L_2-function $x(t)$ then has a Hilbert transform (generalized function) which corresponds to the Hilbert transform of $x(t)$. The uniqueness which Hilbert transform pairs enjoy in L_2 is lost when going to generalized functions. This and some other properties of Hilbert transforms of generalized functions are discussed.

1. INTRODUCTION

Hilbert transforms occur naturally in the representation of causal functions in $L_2(-\infty,\infty)$ in terms of their Fourier spectra. By causal function we mean a function $x(t)$ which vanishes for $t < 0$. Denoting the Fourier transform of the causal function $x(t)$ in L_2 by[*]

$$\mathfrak{F}_x(f) = \int x(t)e^{-j2\pi ft}\,dt, \tag{1}$$

then the real and imaginary parts of $\mathfrak{F}_x(f)$ are a Hilbert transform pair; that is,

$$\left.\begin{aligned} \mathrm{Re}[\mathfrak{F}_x(f)] &= \frac{1}{\pi}\int \frac{\mathrm{Im}[\mathfrak{F}_x(\nu)]}{f-\nu}\,d\nu \equiv \mathcal{H}\mathrm{Im}[\mathfrak{F}_x(f)], \\ \mathrm{Im}[\mathfrak{F}_x(f)] &= -\frac{1}{\pi}\int \frac{\mathrm{Re}[\mathfrak{F}_x(\nu)]}{f-\nu}\,d\nu \equiv -\mathcal{H}\mathrm{Re}[\mathfrak{F}_x(f)], \end{aligned}\right\} \tag{2}$$

where the integrals are to be considered as Cauchy principal values.

The theory of generalized functions has extended the Fourier transformation in a rigorous fashion to a larger class of functions, in fact, to all functions that are likely to arise in the representation of signals. It is therefore desirable to extend also the notion of Hilbert transforms in such a way that it applies within the framework of generalized functions. This is done in the present paper. Some properties of Hilbert transforms of generalized functions are also examined.

---

[*] Limits of integration are $-\infty,+\infty$ where no limits are indicated.

A very readable introduction to the theory of generalized functions ("g.f.'s") has been presented by M.J. Lighthill. A brief summary of the essentials needed for the development to follow is given in the next two sections.

2. GENERALIZED FUNCTIONS

We begin with a class $\mathcal{S}$ of (complex-valued) functions of a real variable which are most well-behaved: they are differentiable an arbitrary number of times and they fall off more rapidly than any power of the independent variable as this variable becomes large in magnitude. Such functions we will denote $\hat{x}(t)$, $\hat{y}(t)$, etc. At the heart of the theory of g.f.'s is the ability to make an appropriate transition from this class of very well-behaved functions to an extremely wide class of functions $\mathcal{S}'$ without abandoning the nice properties of the well-behaved functions in $\mathcal{S}$.

The means for making the desired transition are so-called "weakly-convergent" (w.c.) sequences in $\mathcal{S}$. The sequence $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n, \ldots$ of functions in $\mathcal{S}$, is a "weakly convergent" sequence if, for every function $\hat{y}(t)$ also in $\mathcal{S}$, the sequence of integrals

$$\left\{ \int \hat{x}_n(t)\, \hat{y}(t)\, dt \right\}_{n = 1,2,\ldots} \tag{3}$$

converges. The limit of such a sequence of integrals is symbolically written

$$\int \overline{x(t)}\hat{y}(t)\, dt, \tag{4}$$

where $\overline{x(t)}$ is called the weak limit of the sequence $\hat{x}_1(t), \hat{x}_2(t), \ldots \hat{x}_n(t), \ldots$ Then $\mathcal{S}'$ is the class of all such weak limits of w.c. sequences in $\mathcal{S}$. The "Dirac delta function", $\overline{\delta(t)}$, is a member of $\mathcal{S}'$, for instance; it assigns to the expression (4) the value $\hat{y}(0)$.

The elements of $\mathcal{S}'$ are not functions, in the sense of assigning a value to every point on the real line. Many members of the class $\mathcal{S}'$, however, do correspond to ordinary functions, or portions thereof, in the following way: We say that the g.f. $\overline{x(t)}$ corresponds to the ordinary function $h(t)$ [denoted $\overline{x(t)} \sim h(t)$] on an interval $a < t < b$, if for each $\hat{y}$ in $\mathcal{S}$ which vanishes outside (a,b),

$$\int \overline{x(t)}\hat{y}(t)\, dt = \int h(t)\hat{y}(t)\, dt. \tag{5}$$

By an ordinary function we mean here a function which is absolutely Riemann-integrable over any finite interval. Thus, the g.f. $\overline{\delta(t)}$ corresponds to the ordinary function 0 (i.e., it vanishes) on $(-\infty,0)$ and $(0,\infty)$. It can be shown that there is a g.f. corresponding to every ordinary function $h(t)$ which grows no faster than a power of t, as $|t| \rightarrow \infty$.

3. THE FOURIER TRANSFORM OF A G.F.

One of the very desirable features of the limits which we call g.f.'s is that with each can be associated another g.f. which is its "Fourier transform" (F.T.). To see what is meant when we speak of the F.T. of a g.f., consider first the F.T.'s of functions in $\mathcal{S}$. For any function $\hat{x}(t) \in \mathcal{S}$, the integral (1) converges and is called the F.T. of $\hat{x}(t)$. There is a unique $\mathcal{F}_{\hat{x}}(f) \in \mathcal{S}$ for each $\hat{x}(t) \in \mathcal{S}$. (The use of different letters for the argument, f and t, is for engineering appeal; both $\hat{x}$ and $\mathcal{F}_{\hat{x}}$ are complex functions of a real variable, in $\mathcal{S}$.)

Consider now any w.c. sequence in $\mathcal{S}$, $\{\hat{x}_n\} = \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n, \ldots$, whose weak limit is the g.f. $\overline{x}$; we write $\{\hat{x}_n\} \xrightarrow{W} \overline{x}$. Every member of the sequence $\{x_n\}$ has a F.T. in $\mathcal{S}$ and, furthermore, it turns out that the sequence of F.T.'s, $\{\mathcal{F}_{\hat{x}_n}\}$ is also w.c. Its weak limit is called the F.T. of $\overline{x}$, denoted $\mathcal{F}_{\overline{x}}$. This definition is consistent with the ordinary meaning of F.T.: If the g.f. $\overline{x(t)}$ corresponds on $(-\infty,\infty)$ to an ordinary function $h(t)$ which has a F.T. $\mathcal{F}_h(f)$, then $\mathcal{F}_{\overline{x}}(f) \sim \mathcal{F}_h(f)$ on $(-\infty,\infty)$. For example, if $\{\hat{x}_n(t)\} \xrightarrow{W} \overline{\delta(t)}$, then $\{\mathcal{F}_{\hat{x}_n}(t)\} \xrightarrow{W} \overline{1}$, where $\overline{1} \sim$ the constant function 1 on $(-\infty,\infty)$.

In this sense then every g.f. has a F.T. Other properties of the very well-behaved functions in $\mathcal{S}$ carry over via weak convergence to $\mathcal{S}'$, the most spectacular one being differentiability an arbitrary number of times.

4. THE G.F. $\overline{\text{sgn } t}$

As an example we consider the ordinary function

$$\text{sgn } t = \begin{Bmatrix} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{Bmatrix},$$

to which must correspond, as stated in Sec. 2, a g.f. $\overline{\text{sgn } t}$. Thus for some sequence $\{\hat{x}_n\}$ in $\mathcal{S}$,

$$\lim_{n\rightarrow\infty} \int \hat{x}_n(t)\, \hat{y}(t)\, dt = \int \overline{\text{sgn } t}\, \hat{y}(t)\, dt,$$

for every $\hat{y}(t) \in \mathcal{S}$. Then we have, using a prime to denote the derivative,

$$\int \overline{\text{sgn}'\ t}\ \hat{y}(t)\ dt = \lim_{n\to\infty} \int \hat{x}'_n(t)\ \hat{y}(t)\ dt = -\lim_{n\to\infty} \int \hat{x}_n(t)\ \hat{y}'(t)\ dt$$

$$= -\int \overline{\text{sgn}\ t}\ \hat{y}'(t)\ dt = -\int \text{sgn}\ t\ \hat{y}'(t)\ dt = \left(\int_{-\infty}^{0} - \int_{0}^{\infty}\right) \hat{y}'(t)\ dt = 2y(0);$$

so that (see Sec. 2)

$$\int \overline{\text{sgn}'t}\ \hat{y}(t)\ dt = 2\int \overline{\delta(t)}\ \hat{y}(t)\ dt, \text{ for all } \hat{y} \in \mathcal{S};$$

i.e., $\overline{\text{sgn}'t} = 2\overline{\delta(t)} = \overline{2\delta(t)}$.

We now wish to find the F.T. of $\overline{\text{sgn}\ t}$; that is, we know there exists an element $\mathcal{F}_{\overline{\text{sgn}}} \in \mathcal{S}'$, but we want a more meaningful description of it. For the x_n's as above,

$$\int \mathcal{F}_{\overline{\text{sgn}}}(f)\ \hat{y}(f)\ df = \lim_{n\to\infty} \int \mathcal{F}_{\hat{x}_n}(f)\ \hat{y}(f)\ df$$

$$= \lim_{n\to\infty} \int \frac{2\pi jf\mathcal{F}_{\hat{x}_n}(f)}{2\pi jf}\ \hat{y}(f)\ df = \lim_{n\to\infty} \int \frac{\mathcal{F}_{\hat{x}'_n}(f)}{2\pi jf}\ \hat{y}(f)\ df. \tag{6}$$

Since $\{\mathcal{F}_{\hat{x}'_n}(f)\} \xrightarrow{w} 2\ \mathcal{F}_{\overline{\delta}}(f) = 2.\overline{1} = \overline{2}$, it seems appropriate, in view of (6), to write $\mathcal{F}_{\overline{\text{sgn}}}(f) = \overline{\frac{1}{j\pi f}}$. Now, the function $\frac{\hat{y}(f)}{2\pi jf}$ is in $\mathcal{S}$ only if $\hat{y}(f)$ vanishes at least at the origin. For any such $\hat{y}(f) \in \mathcal{S}$, the limit (6) then equals the ordinary integral

$\int 2 \cdot \frac{\hat{y}(f)}{2\pi jf}\ df = \int \frac{1}{j\pi f}\ \hat{y}(f)\ df$; i.e., the F.T. of $\overline{\text{sgn}\ t}$, which we have denoted $\overline{\frac{1}{j\pi f}}$, in fact corresponds to $\frac{1}{j\pi f}$ on $(-\infty,0)$ and $(0,\infty)$.

5. PRINCIPAL VALUE OF AN ORDINARY INTEGRAL

Next we want to show that the g.f. $\overline{\frac{1}{j\pi f}}$ defines the Cauchy Principal Value (P.V.) of an ordinary integral, namely, that

$$\int \overline{\frac{1}{j\pi f}}\ \hat{y}(f)\ df = \underline{\underline{\text{P.V.}}} \int \frac{1}{j\pi f}\ \hat{y}(f)\ df, \tag{7}$$

for any $\hat{y}(f) \in \mathcal{S}$. First note that the left hand side, being the limit of a convergent sequence of integrals as in (3), must exist for all $\hat{y} \in \mathcal{S}$. Therefore the right-hand side must be defined in such a way as to give this integral a finite value. To see that the correct finite value for the right side is the P.V., write $\hat{y}(f)$ in terms of its even and odd components (which are also in $\mathcal{S}$),

$$\hat{y}(f) = \hat{y}_e(f) + \hat{y}_o(f),$$

and note that $\frac{1}{j\pi f}$ is an odd function. Now,

$$\text{P.V.}\int \frac{1}{j\pi f}\,[\hat{y}_e(f) + \hat{y}_o(f)]\,df = \lim_{\epsilon\to 0}\left(\int_{-\infty}^{-\epsilon} + \int_{\epsilon}^{\infty}\right)\frac{1}{j\pi f}\,[\hat{y}_e(f) + \hat{y}_o(f)]\,df$$

$$= \lim_{\epsilon\to 0}\left(\int_{-\infty}^{-\epsilon} + \int_{\epsilon}^{\infty}\right)\frac{1}{j\pi f}\,\hat{y}_o(f)\,df = \int \frac{1}{j\pi f}\,\hat{y}_o(f)\,df; \tag{8}$$

i.e., taking the P.V. amounts to neglecting the odd part of the integrand. Similarly, for the left side of (7) we have

$$\int \overline{\frac{1}{j\pi f}}\,[\hat{y}_e(f) + \hat{y}_o(f)]\,df = \int \overline{\frac{1}{j\pi f}}\,\hat{y}_e(f)\,df + \int \overline{\frac{1}{j\pi f}}\,\hat{y}_o(f)\,df,$$

where the first term on the right equals, in terms of the $\hat{x}_n$'s of Sec. 4,

$$\int \mathcal{F}_{\overline{\text{sgn}}}(f)\hat{y}_e(f)\,df = \lim_{n\to\infty}\int \mathcal{F}_{\hat{x}_n}(f)\hat{y}_e(f)\,df = \lim_{n\to\infty}\int \hat{x}_n(t)\,\hat{w}_e(-t)\,dt$$

(by Parseval's relation, where $\hat{w}_e$ is the inverse F.T. of $\hat{y}_e$ and is even)

$$= \int \text{sgn}(t)\,\hat{w}_e(-t)\,dt = 0.$$

Thus (7) is verified by observing that

$$\int \overline{\frac{1}{j\pi f}}\,\hat{y}(f)\,df = \int \overline{\frac{1}{j\pi f}}\,\hat{y}_o(f)\,df = \int \frac{1}{j\pi f}\,\hat{y}_o(f)\,df.$$

The last equality follows from Sec. 4, since $\hat{y}_o(f)$ vanishes at the origin.

6. THE HILBERT TRANSFORM OF A FUNCTION IN $\mathcal{S}$

Only a slight modification of (7) is needed to give the Hilbert transform (H.T.) of a function in $\mathcal{S}$. For $\hat{y}(f) \in \mathcal{S}$, $\hat{y}(-f)$ is in $\mathcal{S}$ also, and so is $\hat{y}(f_o - f)$, for all real shifts f_o. Substituting in (7), we see that the Hilbert Transform of $\hat{y}$, evaluated at f_o, exists for all f_o:

$$\mathcal{H}\hat{y}(f_o) = \frac{1}{\pi}\,\text{P.V.}\int \frac{1}{f}\,\hat{y}(f_o - f)\,df = j\int \overline{\frac{1}{j\pi f}}\,\hat{y}(f_o - f)\,df$$

so that $\mathcal{H}\hat{y}$ is an ordinary function, for every $\hat{y} \in \mathcal{S}$. The functions $\mathcal{H}\hat{y}$ can be characterized in greater detail. We denote by $\mathcal{S}_o$ the subclass of $\mathcal{S}$ of all those functions which vanish at the origin, together with all their derivatives.

Theorem: For $\hat{y}(t) \in \mathcal{S}$, if $\mathcal{F}_{\hat{y}} \in \mathcal{S}_o$, then $\mathcal{H}\hat{y}(t) \in \mathcal{S}$ also; if $\mathcal{F}_{\hat{y}} \notin \mathcal{S}_o$, $\mathcal{H}\hat{y}(t) \notin \mathcal{S}$ but is an infinitely differentiable function.

Proof:

a) $\mathcal{F}_{\hat{y}} \in \mathcal{S}_o$. $\mathcal{F}_{\hat{y}}(f)$ can be broken up into two components, vanishing for negative and positive f, respectively, which are the F.T.'s of an "analytic signal", $\hat{z}(t)$ and the conjugate of an analytic signal, $\hat{w}(t)$, respectively, both in $\mathcal{S}$. Then $\mathcal{H}\hat{y} = \mathcal{H}[\hat{z} + \hat{w}] = (-j\hat{z} + j\hat{w}) \in \mathcal{S}$.

b) $\mathcal{F}_{\hat{y}} \notin \mathcal{S}_o$. $\mathcal{H}\hat{y}$ cannot be in $\mathcal{S}$, because the F.T. of $\hat{y} + j\mathcal{H}\hat{y}$ must be causal (Sec. 1), whereas to $\mathcal{F}_{\hat{y}}$ must be added a function outside $\mathcal{S}$ to result in a causal function. The $k^{\underline{th}}$ derivative of $\hat{y}$, $\hat{y}^{(k)}$, is in $\mathcal{S}$, for $k = 0,1,2,\ldots$, so that $\mathcal{H}\hat{y}^{(k)}$ exists as an ordinary function. Then from a property of the H.T. (see also Sec. 9), $(\mathcal{H}\hat{y})^{(k)} = \mathcal{H}[\hat{y}^{(k)}]$ so that $\mathcal{H}\hat{y}$ is infinitely differentiable.

It can also be shown that for $\mathcal{F}_{\hat{y}} \notin \mathcal{S}_o$, $\mathcal{H}\hat{y}(t)$ is $O(t^{-1})$ as $|t| \to \infty$.

7. THE HILBERT TRANSFORM OF A G.F.: DISCUSSION

How can weak convergence be applied to transfer the property "every element has a Hilbert transform" from the space $\mathcal{S}$ to the space $\mathcal{S}'$ of g.f.'s? For any given g.f. $\overline{x(t)}$, there are in $\mathcal{S}$ sequences $\{\hat{x}_n(t)\} \xrightarrow{w} \overline{x(t)}$. However, it is not possible now to proceed in the same way as in the definition of the F.T. for g.f.'s (Sec. 3): A sequence $\{\mathcal{H}\hat{x}_n(t)\}$ in general is made up not of functions in $\mathcal{S}$, but of ordinary functions, as was shown in the previous Section. To every $\mathcal{H}\hat{x}_n(t)$ in such a sequence, of course, corresponds a g.f. $\overline{\mathcal{H}\hat{x}_n(t)}$, but there is no guarantee that the sequence $\{\overline{\mathcal{H}\hat{x}_n(t)}\}$ in $\mathcal{S}'$ is w.c.

Instead, let us look at the F.T. of the sequence $\{j\mathcal{H}\hat{x}_n(t)\}$. This is obtained directly from $\{\hat{x}_n(t)\}$ as $\{\text{sgn } f\, \mathcal{F}_{\hat{x}_n}(f)\}$, which is again a sequence of ordinary functions, in general. We do know that $\{\mathcal{F}_{\hat{x}_n}(f)\}$ is in $\mathcal{S}$ and w.c., since $\{\hat{x}_n(t)\}$ is w.c. But we now need a w.c. sequence of g.f.'s that corresponds in some way to the sequence of ordinary functions $\{\text{sgn } f\, \mathcal{F}_{\hat{x}_n}(f)\}$, and this sequence should converge to (-j) times the F.T. of what we want to call $\mathcal{H}\overline{x(t)}$.

Ideally, then, the nth term of the desired sequence should be the g.f. which corresponds to the ordinary function $\text{sgn } f\, \mathcal{F}_{\hat{x}_n}(f)$, for all n. Regarding convergence of such a sequence we can say the following. Since

$\{\mathcal{F}_{\hat{x}_n}\}$ is w.c., $\{\int_0^\infty \mathcal{F}_{\hat{x}_n}(f)\,\hat{y}(f)\,df\}$ converges for all those $\hat{y}(f) \in \mathcal{S}$ which vanish outside $(0,\infty)$, and $\{\int_\infty^0 \mathcal{F}_{\hat{x}_n}(f)\,\hat{y}(f)\,df\}$ converges for all those $\hat{y}(f) \in \mathcal{S}$ which vanish outside $(-\infty,0)$, so that $\{\int \operatorname{sgn} f\, \mathcal{F}_{\hat{x}_n}(f)\,\hat{y}(f)\,df\}$ converges for all $\hat{y}(f) \in \mathcal{S}_0$.

Therefore, we will require our sequence of g.f.'s to be made up of terms $\overline{g_n(f)}$ which correspond to the ordinary function $\operatorname{sgn} f\, \mathcal{F}_{\hat{x}_n}(f)$ on $(-\infty,0)$ and $(0,\infty)$, for all n. But given such a $\overline{g_n(f)}$, we see that $\overline{g_n(f)} + \overline{\sum_i a_i \delta^{(i)}(f)}$ also satisfies the requirement (where the a_i's are constant multipliers); i.e., there is an infinity of functions $\overline{g_n}$ for each n. Therefore we need as additional requirement that $\{\overline{g_n(f)}\}$ w.c. in $\mathcal{S}'$. This, however, still permits adding to $\{\overline{g_n(f)}\}$ a w.c. sequence of the form $\{\overline{\sum_i a_i \delta^{(i)}(f)}\}$, giving a new w.c. sequence convergent to a different g.f. Thus the F.T. of $j\,\mathcal{H}\overline{x(t)}$ is not unique but is defined only to within a sum of singularity functions at the origin; so that $\mathcal{H}x(f)$ is likewise not unique but is defined only to within an arbitrary polynomial.

A sequence which converges directly to $\mathcal{H}\overline{x(t)}$ is obtained by taking j times the inverse F.T. of a sequence as specified above, since $\{\mathcal{F}_{\overline{g_n}}\}$ w.c. implies $\{\overline{g_n}\}$ is w.c.

8. DEFINITION OF THE HILBERT TRANSFORM IN $\mathcal{S}'$

We formalize the discussion of the previous Section by defining the Hilbert transform of a g.f. $\overline{x}$ in terms of suitable w.c. sequences in $\mathcal{S}'$. More precisely, a class of g.f.'s is defined in this way, which will all be called Hilbert transforms of $\overline{x}$ and differ from each other by an additive (g.f.) polynomial term.

Definition. Given a g.f. $\overline{x(t)}$, let $\{\hat{x}_n(t)\}$ be a sequence in $\mathcal{S}$ such that $\{\hat{x}_n(t)\} \xrightarrow{w} \overline{x(t)}$. The Hilbert transform of $\overline{x(t)}$, denoted $\mathcal{H}\overline{x(t)}$, is any g.f. whose F.T. $\overline{g(f)}$ is the limit of a w.c. sequence $\{\overline{g_n(f)}\}$ in $\mathcal{S}'$, where for each n, $\overline{g_n(f)}$ corresponds to $-j \operatorname{sgn} f\, \mathcal{F}_{\hat{x}_n}(f)$ on $(-\infty,0)$ and $(0,\infty)$.

Proof of consistency: Let $x(t)$ be an L_2-function to which corresponds the g.f. $\overline{x(t)}$. Then one and only one L_2-function, namely $\mathcal{H}x(t)$, corresponds to $\mathcal{H}\overline{x(t)}$ (that is, to one of the g.f.'s which are Hilbert transforms of $\overline{x(t)}$.)

Let $\{\hat{x}_n(t)\} \xrightarrow{w} \overline{x(t)} \sim x(t) \in L_2$, so that $\{\mathcal{F}_{\hat{x}_n}(f)\} \xrightarrow{w} \mathcal{F}_{\overline{x}}(f) \sim \mathcal{F}_x(f) \in L_2$. Also in L_2 is the F.T. of $\mathcal{H}x(t) = -j \operatorname{sgn} f\, \mathcal{F}_x(f) \sim \overline{-j \operatorname{sgn} f\, \mathcal{F}_x(f)}$. By definition of $\{\overline{g_n(f)}\}$, its limit $\overline{g(f)} \sim \overline{-j \operatorname{sgn} f\, \mathcal{F}_x(f)}$ on $(-\infty,0)$ and $(0,\infty)$. Therefore, if there is an L_2-function $\sim \overline{g(f)}$, it must be $-j \operatorname{sgn} f\, \mathcal{F}_x(f)$. It is also clear that at least one of the sequences $\{\overline{g_n(f)}\} \xrightarrow{w} \overline{-j \operatorname{sgn} f\, \mathcal{F}_x(f)}$.

9. SOME PROPERTIES OF THE HILBERT TRANSFORMS IN $\mathcal{S}'$

a) Repeated application of the H.T.

An important feature of the definition of the previous Section, as has already been pointed out, is the lack of uniqueness of the H.T. in $\mathcal{S}'$. If P(t) stands for "arbitrary polynomial", we have $\mathcal{H}\overline{x(t)} = \mathcal{H}\overline{x(t)} + \overline{P(t)}$. The advantage of this is that Hilbert transformation remains a reversible operation, that is, $-\mathcal{H}$ acts as the inverse of $\mathcal{H}$:

$$-\mathcal{H}[\mathcal{H}\overline{x(t)}] = \overline{x(t)} \text{ [within } \overline{P(t)}\text{]}.$$

Thus, for the ordinary function x(t) = t, we have $\mathcal{H}\overline{x(t)} = \mathcal{H}\overline{t} = \overline{P(t)}$; $-\mathcal{H}\overline{P(t)} = \overline{P(t)}$, which includes the result $-\mathcal{H}[\mathcal{H}\overline{t}] = \overline{t} \sim t$.

b) $\mathcal{H}\overline{x'(t)} = [\mathcal{H}\overline{x(t)}]'$

For $\overline{x(t)} \in \mathcal{S}'$, let $\{\hat{x}_n(t)\} \xrightarrow{W} \overline{x(t)}$, and $\{\overline{g_n(f)}\} \xrightarrow{W} \mathcal{F}_{\mathcal{H}\overline{x}}$, where

$$\overline{g_n(f)} \sim -j \operatorname{sgn} f\, \mathcal{F}_{\hat{x}_n}(f) \text{ on } (-\infty,0),\ (0,\infty).$$

Then $\{\hat{x}'_n(t)\} \xrightarrow{W} \overline{x'(t)} \in \mathcal{S}'$. Next let $\{\overline{h_n(f)}\} \xrightarrow{W} \mathcal{F}_{\mathcal{H}\overline{x'}}$, where

$$\overline{h_n(f)} \sim -j \operatorname{sgn} f\, \mathcal{F}_{\hat{x}'_n}(f) = -j \operatorname{sgn} f\,(j2\pi f)\mathcal{F}_{\hat{x}_n}(f),$$

so that $\overline{h_n(f)} \sim \overline{j2\pi f\, g_n(f)}$ on $(-\infty,0)$, $(0,\infty)$. But $\{\overline{j2\pi f g_n(f)}\} \xrightarrow{W} \mathcal{F}_{(\mathcal{H}\overline{x})'}(f)$, so that $\mathcal{F}_{(\mathcal{H}\overline{x})'} \sim \mathcal{F}_{\mathcal{H}\overline{x'}}$ on $(-\infty,0)$, $(0,\infty)$, and $\mathcal{H}\overline{x'} = (\mathcal{H}\overline{x})' + \overline{P}$.

As an example, consider $\mathcal{H}\overline{\delta(t)} = \overline{\frac{1}{\pi t}}$. This gives, again within P(t):

c) $\mathcal{H}\overline{\delta^{(n)}(t)} = \overline{(\frac{1}{\pi t})^{(n)}}$, which corresponds to $\frac{1}{\pi}\,\frac{(-n)!}{t^{n+1}}$ on $(-\infty,0)$, $(0,\infty)$.

Finally, let us see how the "preservation of inner product" property of the H.T. on L_2 extends to $\mathcal{S}'$. Here we have:

d) $\int \mathcal{H}\overline{x(t)}\ \mathcal{H}\hat{y}(t)\,dt = \int \overline{x(t)}\ \hat{y}(t)\,dt$ for $\hat{y}$ such that $\mathcal{F}_{\hat{y}} \in \mathcal{S}_o$. This is easily verified: For $\mathcal{F}_{\hat{y}} \in \mathcal{S}_o$, $\mathcal{H}\hat{y} \in \mathcal{S}$ (Sec. 6), so that the left-hand side is meaningful and equals

$$\int \mathcal{F}_{\mathcal{H}\overline{x}}(f)\,\mathcal{F}_{\mathcal{H}\hat{y}}(-f)\,df = \lim_{n\to\infty}\int -j \operatorname{sgn} f\,\mathcal{F}_{\hat{x}_n}(f)\,[-j \operatorname{sgn}(-f)]\mathcal{F}_{\hat{y}}(-f)\,df$$

$$= \lim_{n\to\infty}\int \mathcal{F}_{\hat{x}_n}(f)\,\mathcal{F}_{\hat{y}}(-f)\,df = \int \mathcal{F}_{\overline{x}}(f)\mathcal{F}_{\hat{y}}(f)\,df = \int \overline{x(t)}\hat{y}(t)\,dt.$$

10. CONCLUSION

We have seen how a Hilbert Transform can be defined for all g.f.'s in $\mathcal{S}'$, consistent with the ordinary definition of the H.T. This permits, for instance, analytic signal representation within the framework of g.f.'s. The most important properties of the H.T., namely its relation to the Fourier Transform, and the effect of repeated application of the H.T., carry over to g.f.'s. The multiplicity of representation arising from the arbitrary additive polynomial should not cause any problems in applications because undesired polynomial terms can usually be identified by inspection.

REFERENCES

1. M.J. Lighthill, Introduction to Fourier Analysis and Generalized Functions, Cambridge University Press, Cambridge, 1962. Chapter 2.

2. A. Papoulis, The Fourier Integral and Its Applications, McGraw-Hill Book Co., New York, 1962. Chapter 10.

3. A.H. Zemanian, Distribution Theory and Transform Analysis, McGraw-Hill Book Co., New York, 1965. Chapters 1 and 4.

A CONJECTURE ON THE EXISTENCE OF A QUADRATIC LIAPUNOV FUNCTION FOR LINEAR TIME-INVARIANT FEEDBACK SYSTEMS

KUMPATI S. NARENDRA
Dunham Laboratory
Department of Engineering and Applied Science
Yale University
New Haven, Connecticut

CHARLES P. NEUMAN
Division of Engineering and Applied Physics
Harvard University
Cambridge, Massachusetts

In recent years there has been considerable interest in the absolute stability of nonlinear feedback systems and numerous papers have appeared in the literature using either a functional analysis approach or the direct method of Liapunov. The contributions of Lur'e, Popov, Yakubovich, Kalman, and Lefschetz are well known in this area. Using the Liapunov approach conditions are generally found for the existence of a particular form of the Liapunov function which is then interpreted in terms of a frequency condition. In this short paper a conjecture is made that if a time-invariant feedback system (Fig. 1) is asymptotically stable for every constant feedback gain K in the interval $0 < K < \infty$, then a Liapunov function of the form $X^T PX + KX^T MX$ exists where P is a positive definite matrix and Q is a nonnegative definite matrix. The importance of this result arises from the fact that parts of this Liapunov function can be suitably modified to prove the stability of various nonlinear systems.

In this paper only the linear time-invariant case is considered. The extension to various nonlinear situations as well as the interpretation of the results in terms of Aizerman's conjecture will appear elsewhere (Ref. 1).

THE PROBLEM

A single-loop negative feedback system (Fig. 1) consists of a stable linear system G(s) in the forward path and a constant gain K in the feedback path. Assuming that the system is of n^{th} order, it can be represented in the state vector form as

$$\begin{aligned} \dot{X} &= AX + bu \\ \sigma &= h^T X \\ u &= -K\sigma \end{aligned} \qquad (1)$$

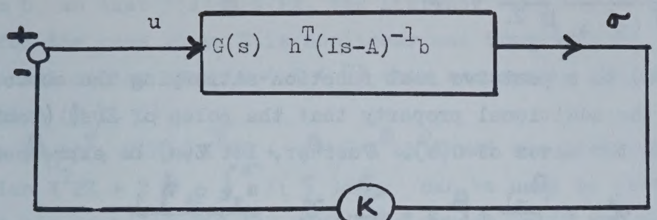


Figure 1

The over-all system of (1) may be represented by the vector equation

$$\dot{X} = \left[A - K\, bh^T\right] X \qquad (2)$$

This homogeneous system is assumed to be asymptotically stable for all values of the feedback gain K in the interval $0 < K < \infty$. It is well known that the stability of a linear time-invariant system can be determined using a Liapunov function of the quadratic form. Since the system (2) is asymptotically stable for every K in the interval $0 < K < \infty$, for any specific value of K a quadratic Liapunov function exists. However, a single Liapunov function has not been determined so far for establishing the asymptotic stability of system (2) for all values of K in the interval. In this paper a conjecture is made regarding the existence of such a Liapunov function.

THE CONJECTURE

If the linear system G(s) represented in Equation (1) is completely controllable, completely observable, and asymptotically stable, and if the over-all system is asymptotically stable in the interval $0 < K < \infty$, then there exists a Liapunov function for the system having the form $X^TPX + KX^TQX$, where the positive definite matrix P and the positive semidefinite matrix Q are independent of the feedback gain and X^TPX is a Liapunov function for the asymptotically stable plant.

The basis for this conjecture may now be outlined: (i) If the system (2) is asymptotically stable for all values of gain K in the interval $0 < K < \infty$, then the Nyquist plot of $G(i\omega)$ does not intersect the negative real axis and $-\pi < \arg G(i\omega) < +\pi$ for all real 'ω.' If G(s) is expressed as $\frac{m(s)}{n(s)}$, a positive real function Z(s) exists such that $G(s) \cdot Z(s)$ is positive real. (One such function Z(s) was suggested by Brockett and Willems (Ref. 2) and has the form

$$Z(s) = s\frac{\prod_i (s^2 + \lambda_i^2)}{\prod_j (s^2 + \mu_j^2)})$$

(ii) Let Z(s) be a positive real function satisfying the above condition and having the additional property that the poles of Z(s) (real or complex) are the zeros of G(s). Further, let Z(s) be expressed as

$$Z(s) = \Big[\alpha_0 + \frac{\beta_{-1}}{s} + \beta_0 s + \sum_i \gamma_i \frac{s + c_i \eta_i}{s + \eta_i} + \sum_j \delta_j \frac{s^2 + \lambda_j s + \lambda_{aj}}{s^2 + \mu_j s + \mu_{aj}} \Big] \tag{3}$$

where α_0, β_{-1}, β_0, γ_i, and δ_j are nonnegative.

(iii) The Meyer-Yacubovich-Kalman lemma (Ref. 3, 4, 5) can be used to show that the frequency domain condition [G($i\omega$) Z($i\omega$) is positive real, where Z($i\omega$) has the form (3),] corresponds to the existence of a positive definite matrix P and a positive semidefinite matrix D and a vector q which simultaneously satisfy the equations

$$\begin{aligned} A^T P + PA &= -2qq^T - D \\ Pb - d &= 2\sqrt{\gamma}\, q \end{aligned} \tag{4}$$

where $d = \alpha_0 h^T + \beta_0 h^T A + \beta_{-1} h^T A^{-1} + \sum_i \gamma_i r_i^T + \sum_j \delta_j s_j^T$ (5)

and the vectors r_i and s_j are suitably chosen (Ref. 6, 7).

(iv) In such a case a Liapunov function of the form

$$V(X) = X^T PX + KX^T \Big[\beta_0 hh^T + \sum_{i=1}^{n} \epsilon_i r_i r_i^T + \sum_j \delta_j s_j s_j^T\Big] X \tag{6}$$

may be shown to exist. Or, $V(X) \quad X^T PX + K X^T MX$

where $M = \beta_0 hh^T + \sum \epsilon_i r_i r_i^T + \sum {}_j s_j s_j^T$.

The existence of the Liapunov function V(X) of the form given in (6) depends entirely on (ii), i.e., whether a Z(s) can always be found which is positive real, has the form $\frac{T(s)}{m(s)}$, and can be expanded in the form (3). The verification of this conjecture would then require the demonstration of the existence of such a function.

Importance of the Conjecture

If the conjecture is proved to be true, the stability of linear and nonlinear systems with a single gain function can be treated in a unified manner. The Liapunov function obtained in the time-invariant case can be used to determine the stability of the feedback system for various nonlinear and time-varying situations as shown below.

a) If $M = 0$ so that $V(X) = X^T PX$, the Liapunov function can also be used for the case where K is nonlinear and time-varying. For the finite sector case ($0 < K < \bar{K}$) this reduces to the circle criterion (Ref. 9).

b) If $M = \beta_0 hh^T$, $V(X) = X^T PX + \beta_0 K X^T hh^T X$ and the Lur'e Liapunov function $X^T PX + 2 \beta_0 \int_0^{\sigma} f(\zeta) d\zeta$ can be used to prove the asymptotic stability of the nonlinear system, with the nonlinear feedback function $F(\sigma) = \sigma K(\sigma)$ lying in the first and third quadrants.

c) If $M = \beta_0 hh^T + \sum_i \gamma_i r_i r_i^T$, Liapunov functions of the form

$$X^T PX + \beta_0 \int_0^{\sigma} f(\zeta) d\zeta + \sum_i \gamma_i \int_0^{r_i^T x} f(\zeta) d\zeta$$

can be used to show the stability of the nonlinear feedback system with monotone or odd monotone nonlinear functions (Ref. 6, 7).

Finally, if the existence of a Liapunov function of the form $X^T PX + K X^T QX$ is proved for the time-invariant case, it can be modified to the form $X^T PX + K(t) X^T MX$ for the time-varying problem where $\dot{K}(t)$ is bounded. When $\dot{K}(t)$ is unbounded, $M \equiv 0$ and the Liapunov function $X^T PX$ yields the circle criterion. When $\dot{K}(t) = 0$ the Nyquist criterion yields the range over which the system is stable. The time-varying Liapunov function $X^T PX + K(t) X^T MX$ will, it is felt, yield a geometrical criterion for the stability of the more practical problem which arises when $\dot{K}(t)$ is bounded.

REFERENCES

1. Kumpati S. Narendra and Charles P. Neuman, "A Quadratic Liapunov Function for the Stability of Linear Time Invariant Feedback Systems," Dunham Laboratory Technical Report CT-6, Yale University.
2. R. W. Brockett and J. L. Willems, "Frequency Domain Stability Criteria--Part I," IEEE Transactions on Automatic Control, AC-10 (1965) pp. 255-261.
3. K. R. Meyer, "Liapunov Functions for the Problem of Lur'e," Proc. Nat. Acad. Sci. U.S.A., 53 (1965) pp. 501-503.
4. R. E. Kalman, "Liapunov Functions for the Problem of Lur'e in Automatic Control," Ibid., 49 (1963), pp. 201-205.
5. V. A. Yakubovich, "The Solution of Certain Matrix Inequalities in Automatic Control Theory," Dokl. Akad. Nauk SSSR, 143 (1962) pp. 1304-1307.
6. K. S. Narendra and C. P. Neuman, "Stability of a Class of Differential Equations with a Single Monotone Nonlinearity," SIAM Journal on Control, 4 (1966), No. 2.

7. M. A. L. Thathachar, M. D. Srinath, and H. K. Ramapriyan, "Solution of a Modified Lur'e Problem," (to be published).

8. K. S. Narendra and R. M. Goldwyn, "Existence of Quadratic Type Liapunov Functions for a Class of Nonlinear Systems," Internat. J. Engrg. Sci., 2 (1964) pp. 367-377.

9. K. S. Narendra and R. M. Goldwyn, "A Geometrical Criterion for the Stability of Certain Nonlinear Nonautonomous Systems," IEEE Trans. of the Circuit Theory Group, CT-11 (1964) pp. 406-408.

THE NETWORK ANALOG AS A HEURISTIC TOOL FOR GENERATING STABILITY CRITERIA*

THOMAS E. STERN
Department of Electrical Engineering
Columbia University
New York, New York 10027

ABSTRACT

A network consisting of a current source, a voltage source and a number of nonlinear time-varying resistive and inductive elements is exhibited. It is shown that the behavior of this network is analogous to that of some fairly general types of feedback systems, and that by properly choosing the terminal characteristics of the network elements, a variety of stability criteria for the analogous feedback systems can be deduced. As an example of the utility of the network analog, a new criterion for absolute stability of nonlinear time-varying feedback systems is derived.

1. INTRODUCTION

We consider feedback systems of the general form shown in Fig. 1, wherein g is the input, w the output, f the error and r may be considered to account for external disturbances as well as any effects of initial conditions on the block $\underline{H}$. We assume that $\underline{H}$ is a linear time-invariant nonanticipative operator defined by[1]

$$(\underline{H}\ f)(t) = \int_0^t h(t-\tau)f(\tau)d\tau \qquad t \geq 0 \tag{1}$$

and $\underline{\varphi}$ is a memoryless operator which may be nonlinear and time-varying, defined by

$$(\underline{\varphi}x)(t) = \varphi[x(t),t] \tag{2}$$

So that the complete system is defined by the equations

$$\begin{aligned} g &= f + w \\ w &= \underline{\varphi}[\underline{H}\ f + r] \end{aligned} \tag{3}$$

Our objective is to determine criteria for absolute stability of this system; that is, to determine under what restrictions on $\underline{H}$, the system will be stable for all operators $\underline{\varphi}$ in a given class.

Popov [1] excited considerable interest in this problem by demonstrating a simple frequency-response criterion which assures absolute stability for a broad class of autonomous feedback systems. His work was considerably generalized by Sandberg [2] and Zames [3] as well as many others. Brockett and Willems [4] showed that by placing certain additional restrictions on $\underline{\varphi}$ one can relax the restrictions on $\underline{H}$ which are required to prove stability. In effect, they showed

---

*This work was partially supported by the National Science Foundation under grant 2789 and by the U.S. Navy under contract NONR4259(04). Part of the work was performed while the author was a Fulbright Research Scholar at the Faculte des Sciences of the Universite de Paris, France.

[1] A single symbol x is used to represent either a single quantity or a complete time function defined on the interval $[0,\infty)$. The context will make it clear which is meant. An underlined letter $\underline{F}$ represents an operator. Thus, $\underline{F}\ x$ represents the complete time function which is the result of operating on x with $\underline{F}$, while $x(t)$ and $(\underline{F}\ x)(t)$ represent the corresponding time functions evaluated at a particular instant t.

that there is a certain trade-off between the restrictions one must place on the linear and nonlinear portions of the system. Most of the aforementioned work was based on the idea of splitting the "loop gain" operator $\underline{\varphi}\,\underline{H}$ into a product of two operators $(\underline{\varphi}\,\underline{Y})(\underline{L})$ each having certain "passivity" properties,[2] and then using certain inequalities based on energy-like quantities, to show that the system must be stable. The fact that quantities analogous to energy play a fundamental role in proving these criteria suggests that an electrical network analog would be a useful heuristic tool for discovering and proving such stability criteria.

In Section 2 we shall state a (new) specific example of a "Popov-like" stability theorem, and in Section 3 a network analog will be exhibited which suggests the lines of the proof. In Section 4 we study the energy balance in the network analog to complete the proof of the theorem, and in conclusion we suggest further applications of the network analog. The primary purpose of the paper is not so much to state new results, but to point out new lines of reasoning. We follow, for the most part, the general formulation and symbolism of Sandberg. The theorem stated in the next section can be considered as a combination and generalization of certain results of Sandberg, Zames, and Brockett and Willems.[3]

2. A STABILITY THEOREM

Let $\mathcal{L}_2$ denote the class of all functions x, measurable on $[0,\infty)$ for which $\int_0^\infty x^2(t)dt$ exists and is finite.

Let $\mathcal{E}$ denote the class of all functions x, measurable on $[0,\infty)$ for which $\int_0^T x^2(t)dt$ exists and is finite for each $T \in (0,\infty)$.

Let

$$\| x \|_T \triangleq \left[\int_0^T x^2(t)dt \right]^{\frac{1}{2}}, \qquad T > 0$$

$$\| x \| \triangleq \left[\int_0^\infty x^2(t)dt \right]^{\frac{1}{2}}$$

Let $N(\alpha,\beta,\gamma)$ denote the class of all operators $\underline{\varphi}$ defined by (2) for which

a) the inverse operator $\underline{\varphi}^{-1}$ exists

b) the function $\varphi(x,t)$ and $\frac{\partial\varphi}{\partial t}$ are everywhere continuous in (x,t)

c) $0 < \alpha x^2 \leq x\varphi(x,t) \leq \beta x^2 \quad \forall\ (x,t), \quad x \neq 0$ (4)

d) $\frac{\partial\Phi}{\partial t} \leq \gamma\varphi^2 \quad \forall\,(x,t)$, where $\Phi(x,t) \triangleq \int_0^x \varphi(u,t)du, \quad \gamma \geq 0$ (5)

Let RL_0 denote the class of all functions $Z(s)$ of the form $s+\omega_0$, $\omega_0 > 0$.

Let RL_m denote the class of all rational functions $Z(s)$ of the form

$$Z(s) = s + \omega_0 + \sum_{k=1}^{m} \frac{c_k\omega_k s}{s + \omega_k} \tag{6}$$

$$c_k > 0, \ k = 1, 2, \ldots, m$$

$$\omega_k > 0, \ k = 0, 1, \ldots, m$$

(Note that the members of RL_m are (normalized) driving point impedances of lumped passive RL networks.)

---

[2] Other interesting absolute stability criteria, for example the "critical disk" criteria, fit into a somewhat different category and will not be discussed here.

[3] See Sandberg [2] Corollary 1, Zames [3] Theorem 6, and Brockett and Willems [4] Part I, Theorem 3.

Let $\dot{h}(t)$ exist on $[0,\infty)$ and let

$$H(j\omega) = \int_0^\infty h(t)\epsilon^{-j\omega t}dt$$

be the system function associated with $\underline{H}$, and let

$$\sup_{0\leq \omega<\infty} |H(j\omega)| = k < \infty \tag{7a}$$

$$\sup_{0< \omega<\infty} |j\omega H(j\omega)| = k' < \infty \tag{7b}$$

Theorem. In the system defined by (3) let $\underline{\varphi} \in N(\alpha,\beta,\gamma)$, $f \in \mathcal{E}$, g, r, $\dot{r} \in \mathcal{L}_2$. If for some nonnegative integer m there exists a $Z(s) \in RL_m$ and a real number ρ such that

$$\inf_{0\leq\omega<\infty} \left\{\mathrm{Re}[Z(j\omega)H(j\omega)]\right\} + \frac{\omega_o}{\beta} - \gamma\left[1+\left(\frac{\beta}{\alpha}\right)^2 \sum_{k=1}^{m} c_k\right] = \rho > 0 \tag{8}$$

then

$$\|w\| \leq \frac{1}{\rho}\left[\|\dot{r}\| + a\|r\| + b\|g\|\right] + \left\{\frac{\Phi[r(0),0]}{\rho}\right\}^{\frac{1}{2}} \tag{9}$$

where a and b are positive constants dependent upon the form of $\underline{H}$ and $Z(s)$ but independent of r and g.

Note that this result sets an upper bound on the "size" of the output w proportional to the size of the excitations g and r, provided that a certain frequency response condition on $\underline{H}$ is satisfied. It is in this sense that we use the word stability. (In dealing with non-autonomous systems defined by nonlinear functional equations, the more conventional definitions of stability are not particularly useful.) As we shall see, the size of the positive constant ρ has a definite physical interpretation in a network analog of the system, being a measure of the ability of the network to absorb energy; that is, its effective passivity. As ρ approaches zero, (9) shows that the bound on $\|w\|$ becomes arbitrarily large.

In the case $\gamma=0$, which normally occurs only in time-invariant systems, condition (8) reduces to a form similar to Popov's frequency response condition. In most practical cases, the more rapid the time variations in $\underline{\varphi}$, the more difficult it is to satisfy (8).

3. A NETWORK ANALOG

Consider the network of Fig. 2, wherein I and V represent independent current and voltage sources respectively, and the box labelled $\underline{L}$ is a linear time-invariant nonanticipative element defined by

$$e(t) = (\underline{L}\, i_\ell)(t) = \ell_o i_\ell(t) + \int_0^t \ell(t-\tau)i_\ell(\tau)d\tau \qquad t\geq 0 \tag{10}$$

(note that (10) implies that there are no initial conditions on $\underline{L}$). The impedance of $\underline{L}$ is thus

$$L(j\omega) = \ell_o + \int_0^\infty \ell(t)\epsilon^{-j\omega t}dt$$

The box labelled $\underline{\varphi}\ \underline{Y}$ is a (possibly) nonlinear time-varying element defined by

$$i = \underline{\varphi}\ \lambda_o \tag{11a}$$

where $\underline{\varphi}$ is defined in (2), and

$$\lambda_o = \underline{Y}\ e_y + \Lambda \tag{11b}$$

The operator $\underline{Y}$ is linear, and Λ is a time function representing the effects of initial conditions in the box $\underline{\varphi}\ \underline{Y}$.

Combining the above terminal relations with Kirchhoff's laws, we obtain

$$I = i_\ell + i \tag{12a}$$

$$i = \underline{\varphi}[\underline{Y}\,\underline{L}\,i_\ell + \underline{Y}\,V + \Lambda] \qquad (12b)$$

Comparing (12) and (3) and making the identification

$$I \leftrightarrow g, \quad i_\ell \leftrightarrow f, \quad i \leftrightarrow w$$

it can be seen that (12) and (3) will be identical and hence the network will be an appropriate analog of the feedback system provided that we choose

$$\underline{Y}\,\underline{L} = \underline{H} \qquad (13)$$

$$\underline{Y}\,V + \Lambda = r \qquad (14)$$

Note that the excitations in the feedback system are represented by current and voltage sources (and the response to initial conditions) in the analogous network, the output of the feedback system is represented by the current i, and the loop gain operator $\underline{\varphi}\,\underline{H}$ in the feedback system is represented by the combination of the two elements $\underline{\varphi}\,\underline{Y}$ and $\underline{L}$. The way in which these network elements are selected is determined by the way in which we choose to split up the operator $\underline{\varphi}\,\underline{H}$. Different choices of these two elements lead to different stability criteria.

To prove the theorem stated above we shall need an element $\underline{\varphi}\,\underline{Y}$ for which

$$(\underline{Y}\,e_y)(t) = \int_0^t y(t-\tau)e_y(\tau)d\tau \qquad t \geq 0 \qquad (15)$$

and $Z(s) \in RL_m$ for some nonnegative integer m, where

$$Z(s) = 1/Y(s), \quad Y(s) = \int_0^t y(t)e^{-st}dt \qquad (16)$$

That is, $\underline{Y}$ will represent the admittance operator of a lumped passive RL network. We shall show below that it is possible to synthesize a network having the desired terminal relation $\underline{\varphi}\,\underline{Y}$ using a canonical structure identical to that used in synthesis of linear constant RL driving point functions, but with the linear elements replaced by nonlinear time-varying elements.

The two conditions (13) and (14) plus the requirements (15) and (16) on $\underline{\varphi}\,\underline{Y}$ are sufficient to determine the specific form of $\underline{L}$, V, the terminal relations of the resistive and inductive elements in $\underline{\varphi}\,\underline{Y}$ and the required initial flux-linkages in the inductors. We now treat each of these elements separately.

$\underline{L}$: From (13) we have

$$L(s) = Z(s)\,H(s) \qquad (17a)$$

from which, using (1), (10), (6), we obtain

$$\ell_0 = h(0+)$$

$$\ell(t) = \dot{h}(t) + \omega_0 h(t) + \sum_{k=1}^{m} c_k \omega_k \left[h(t) - \omega_k \int_0^t \epsilon^{-\omega_k(t-\tau)} h(\tau)d\tau\right] \qquad (17b)$$

which is a complete time-domain characterization of $\underline{L}$.

$\underline{\varphi}\,\underline{Y}$: This element can be synthesized using the structure shown in Fig. 3 (a well-known Foster-like canonical form). Letting λ_k represent the flux-linkage in L_k, we use inductive elements defined by the terminal relations

$$i = \underline{\varphi}\,\lambda_0 \qquad (18a)$$

for L_0, and

$$i_{L_k} = \underline{\varphi}(\lambda_k/c_K) \qquad (18b)$$

for $L_1, \ldots, L_m$. Note that in the time-invariant case, these elements are conservative, but when $\underline{\varphi}$ is time-varying they may serve as sources or sinks of energy. The resistive elements shall be defined by

$$e_{R_0} = \omega_0 \underline{\varphi}^{-1} i \qquad (19a)$$

for R_0 and

$$e_{R_k} = c_k\omega_k(\underline{\varphi}^{-1} i - \underline{\varphi}^{-1} i_{L_k}) \tag{19b}$$

for $R_1, \ldots, R_m$. While (19b) may look rather odd, in the case where φ is linear (possibly time-varying) it reduces to the more reasonable form

$$e_{R_k} = c_k\omega_k\, \underline{\varphi}^{-1} i_{R_k} \tag{19c}$$

In either case we shall show below that these resistive elements have very convenient passivity properties. The reader may easily verify that with the above terminal relations the structure of Fig. 3 has a terminal relation of the form (11) with $\underline{Y}$ defined as in (15) and (16).

V: We now must choose V and the initial inductor flux-linkages $\lambda_k(0)$ so that (14) is satisfied. A little algebra shows that an appropriate choice is

$$\lambda_o(0) = r(0) \tag{20a}$$

$$\lambda_k(0) = 0, \quad k=1, 2, \ldots, m \tag{20b}$$

$$V(t) = \dot{r}(t) + \omega_o r(t) + \sum_{k=1}^{m} c_k\omega_k\Big[r(t) - \omega_k \int_0^t \epsilon^{-\omega_k(t-\tau)} r(\tau)d\tau\Big] \tag{20c}$$

(Other choices of initial conditions are possible. However, this one simplifies the development somewhat.) The above terminal relations together with the choice, I = g, for the current source, complete the description of the analog network.

4. ENERGY BALANCE: PROOF OF THEOREM

In this section we shall determine a bound on $\|i\|$ in the network of Fig. 2, which when translated into terms of $\|w\|$ gives us the result (9). The bound will be determined by writing the equations of conservation of energy for the network and then utilizing some well-known inequalities (listed in the appendix). Before writing the energy relations it is convenient to make a Thevenin source transformation converting the network to the form shown in Fig. 4, where the equivalent voltage source V' is defined by $V' = \underline{L}\, I + V$. Now, noting that the net power flow into all of the network elements must be zero at each instant of time and integrating the power flow over an interval, [0,T], we obtain

$$E_{V'} = E_L + \sum_{k=0}^{m} E_k + \sum_{k=0}^{m} W_k \tag{21}$$

where $E_{V'}$ represents the total energy delivered by the source V', E_L represents that absorbed by $\underline{L}$, E_k represents that absorbed by R_k, and W_k represents that absorbed by L_k (all over the interval [0,T]). We now evaluate and obtain bounds on each of these terms.

Using (A4) and (17a),

$$E_L = \int_0^T i(t)(\underline{L}i)(t)dt \ge \rho_\ell \|i\|_T^2 \quad \text{where} \quad \rho_\ell = \inf_{0\le\omega<\infty} \mathrm{Re}[L(j\omega)] \tag{22}$$

and thus for purposes of energy calculations the element $\underline{L}$ can be replaced by a resistance ρ_ℓ in series with a passive element.

In calculating the remaining energy bounds we assume that $\underline{\varphi} \in N(\alpha,\beta,\gamma)$ as stated in the hypothesis of the theorem. Using (18a), (19a), (2) and (4) we have

$$E_o = \int_0^T i(t)e_{R_o}(t)dt \ge \frac{\omega_o}{\beta} \|i\|_T^2 \tag{23}$$

and thus R_o may be replaced by a positive resistance ω_o/β in series with a passive element. Similarly, utilization of (19b), (18a), (18b) and the monotonicity of φ gives

$$E_k = \int_0^T i_{R_k}(t)e_{R_k}(t)dt \ge 0 \qquad k=1, 2, \ldots, m \tag{24}$$

that is each R_k behaves as if it is passive. For L_o we have

$$W_o = \int_o^T i(t)\dot{\lambda}_o(t)dt = \Phi[\lambda_o(T),T] - \Phi[\lambda_o(0),0] - \int_o^T \frac{\partial\Phi}{\partial t}[\lambda_o(t),t]dt \quad (25)$$

where the first two terms represent the difference between magnetic energy stored at time T and at time 0, while the last term represents the energy transferred to the network due to time variations in the inductor. Dropping the energy storage at time T, and using (5) and (18a) we find

$$W_o \geq -\Phi[\lambda_o(0),0] - \gamma \, \|i\|_T^2 \quad (26)$$

indicating that the time variations in L_o may be accounted for by a resistance of value $-\gamma$ in series with a passive element. Similar calculations for the remaining inductors yield

$$W_k \geq -\gamma \, (\tfrac{\beta}{\alpha})^2 c_k \, \|i\|_T^2 \quad (27)$$

where we have used (5), (18b), (4) and (A3). (Recall that the initial energy storage is zero in these elements.)

The above calculations show that for purposes of calculating energy bounds, each network element which is not a source can be replaced by a linear fixed resistance plus a passive element. Thus the equivalent network of Fig. 4 can be replaced by the source V' driving a passive element in series with a linear fixed resistance of value

$$\rho = \rho_\ell + \frac{\omega_o}{\beta} - \gamma\left[1 + (\tfrac{\beta}{\alpha})^2 \sum_{k=1}^{m} c_k\right] \quad (28)$$

where from (17a) and (22)

$$\rho_\ell = \inf_{0 \leq \omega < \infty} \left\{\mathrm{Re}[Z(j\omega)\, H(j\omega)]\right\}$$

The energy balance relation (21) can therefore be replaced by

$$E_{V'} + \Phi[\lambda_o(0),0] \geq \rho \, \|i\|_T^2 \quad (29)$$

which simply states that the total energy absorbed by the equivalent resistance ρ cannot be more than the total energy supplied by the source V' plus the initial energy stored in the inductors.

Relation (29) is the fundamental inequality obtained from the network analog. Comparing (28) with (8), we note that the resistance ρ is the analog of the constant ρ appearing in the theorem. Thus it can be seen that the frequency response hypothesis (8) of the theorem is equivalent to requiring that this resistance be positive. This hypothesis is the key to the remainder of the proof.

To complete the proof we use several inequalities to reduce (29) to the desired form. Applying (A2) and the quadratic formula, and simplifying, we obtain

$$\|i\|_T \leq \frac{\|V'\|_T}{\rho} + \left\{\frac{\Phi[\lambda_o(0),0]}{\rho}\right\}^{\frac{1}{2}} \quad (30)$$

Using (A1) and (A3) on (30), (17a) and (20) the result is

$$\|i\| \leq \frac{1}{\rho}\left\{\|\dot{r}\| + [\omega_o + \sum_{k=1}^{m} c_k\omega_k]\|r\| + [k' + k(\omega_o + \sum_{k=1}^{m} c_k\omega_k)]\|I\|\right\} + \left\{\frac{\Phi[\lambda_o(0),0]}{\rho}\right\}^{\frac{1}{2}} \quad (31)$$

where the subscript T has been dropped since the inequality is valid for all T > 0. Comparing (31) and (9) and recalling the correspondence between network and system variables, it is seen that the proof is complete.

5. CONCLUSIONS

It has been shown that by calculating energy balances in an appropriate network analog, it is possible to deduce a stability criterion for certain types of feedback systems. Since no "natural" energy relations are present in the original feedback system while such relations are obvious in the analog, the network analog offers some conceptual advantages in suggesting the lines of the proof of this criterion.

We have presented but one example of the use of such an analog. By changing the terminal characteristics of the network elements many other criteria can be proved. For example, the monotonicity restriction on φ can be either omitted or replaced by various weaker restrictions. (The requirement that $\underline{\varphi}^{-1}$ exist is unnecessary and was included merely to simplify the exposition.) Also, the operator $\underline{\varphi}$ can be restricted to be either time-invariant or linear time-varying. In all of these cases the statement and proof of the stability criterion remains essentially unchanged except for the frequency response hypothesis (8). As should be expected, this hypothesis becomes more restricted as the conditions on φ are relaxed, and vice-versa. Finally, it is possible to construct analogs of a different structure which lead to other types of stability criteria. For example, multi-port network elements can be used as analogs for certain types of multi-loop feedback systems, and networks containing nonlinear resistances and capacitances can be used to generate stability criteria in which the multipliers Z(s) are RC driving point impedances.

APPENDIX

The following inequalities are used at various points in this paper. We list them without proof.

A1. Let $x,y \in \mathcal{E}$. Then

$$\|x+y\|_T \le \|x\|_T + \|y\|_T \quad \forall \quad T > 0 \qquad \text{(triangle inequality)}$$

A2. Let $x,y \in \mathcal{E}$. Then

$$\int_0^T x(t)y(t)dt \le \|x\|_T \|y\|_T \quad \forall \quad T > 0 \qquad \text{(Schwarz inequality)}$$

A3. Let $x \in \mathcal{E}$ and let

$$y(t) = h_o x(t) + \int_0^t h(t-\tau)x(\tau)d\tau$$

where $\sup\limits_{0\le\omega<\infty} \left|h_o + \int_0^\infty h(t)\epsilon^{-j\omega t}dt\right| = k$

Then $\|y\|_T \le k\|x\|_T \quad \forall \quad T > 0$

A4. Let $x \in \mathcal{E}$ and let

$$y(t) = h_o x(t) + \int_0^t h(t-\tau)x(\tau)d\tau$$

where $\inf\limits_{0\le\omega<\infty} \operatorname{Re}[h_o + \int_0^\infty h(t)\epsilon^{-j\omega t}dt] = \rho$

Then $\int_0^T x(t)y(t)dt \ge \rho \; \|x\|_T^2$

REFERENCES

1. V.M. Popov, "Absolute Stability of Nonlinear Systems of Automatic Control," Avtomatika i Telemekhanika, 22, 961-979, 1961.
2. I.W. Sandberg, "Some Stability Results Related to Those of V.M.Popov," Proceedings of the Symposium on Network Theory, Cranfield, 1965.
3. G. Zames, "On the Stability of Nonlinear Time-varying Feedback Systems," Proceedings of the National Electronics Conference, 20, October 1964, p. 725.
4. R.W. Brockett and J.L. Willems, "Frequency Domain Stability Criteria, Parts I and II," IEEE Trans., AC-10, 255-261, 407-413, 1965.

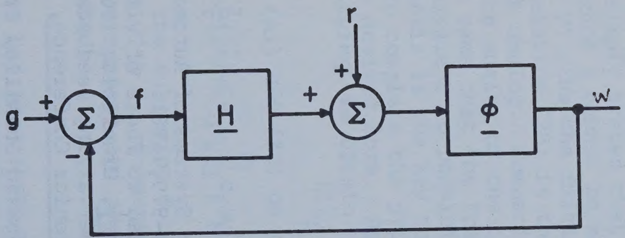


Fig. 1

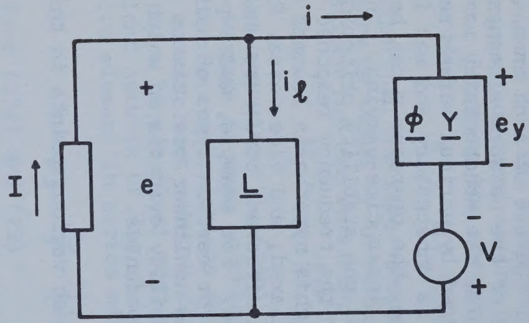

i
I
+
e
−
i_ℓ
L
+
Φ Y
e_y
−
−
V
+

Fig. 2

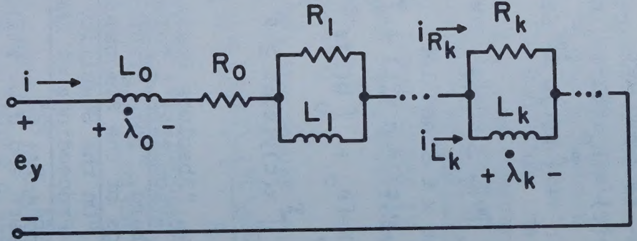

i
+
e_y
−
L_0
+ λ̇_0 −
R_0
R_1
L_1
i_{R_k}
R_k
i_{L_k}
L_k
+ λ̇_k −

Fig. 3

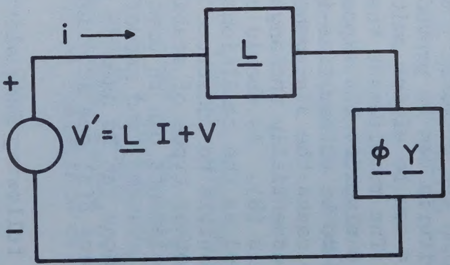


Fig. 4

ON A GENERALIZATION OF THE CIRCLE CRITERION*

MICHAEL GRUBER
Massachusetts Institute of Technology
Cambridge, Massachusetts

JACQUES L. WILLEMS
University of Ghent
Ghent, Belgium

ABSTRACT

It is reasonable to expect that the circle criterion for the stability of time-varying feedback systems can be improved upon when the derivative of the time-varying gain is sufficiently small. In fact it is possible to prove that if a linear feedback system is asymptotically stable for all constant gains $k_1 < k < k_2$ then the system will also be asymptotically stable for all time-varying gains $k_1 < f(t) < k_2$ provided f(t) is sufficiently small. In this paper it is shown that a time-varying feedback system with G(s) in the forward loop and f(t) in the feedback loop will be asymptotically stable if for some Z(s)

(i) $Z(s)[G(s) + \frac{1}{k}]$ and $Z(s-\alpha)$ are positive real

(ii) $0 \leq f(t) < k$

(iii) $\frac{\dot{f}(t)}{f(t)} \leq 2\alpha(1 - \frac{f(t)}{k})$

This result is a generalization of a result presented at the 1964 Allerton Conference by Brockett and Forys in which Z(s) was restricted to be the driving point impedance of either an RL or RC network. For time invariant systems the condition that $Z(s)[G(s) + 1/k_1]$ be positive real for some positive real Z(s) is equivalent to the Nyquist criterion in that it is a necessary and sufficient condition for stability for all k in the range $0 < k < k_1$. Hence if $\dot{f}(t)$ satisfies (iii) the present result shows that the Nyquist criterion also predicts stability for the time-varying system.

PRELIMINARIES

Consider the feedback configuration shown in figure 1. It contains a single time-varying gain in a feedback loop around a linear constant system. It will be assumed that $G(s) = q(s)/p(s)$ where q(s) and p(s) are polynomials in s given by

$$q(s) = q_{n-1}s^{n-1} + q_{n-2}s^{n-2} + \; . \; . \; . \; . + q_1 s + q_o \tag{1}$$

$$p(s) = s^n + p_{n-1}s^{n-1} + . \; . \; . \; . \; . + p_1 s + p_o \tag{2}$$

and the p_i and q_i are real constants any of which may be zero.

---

*This work was supported in part by the National Aeronautics and Space Agency through NASA Research Grant NGR22-009-124 with the M. I. T. Electronics Systems Laboratory.

The original statement of our main result appeared in reference 9. The proof given here follows that of reference 10.

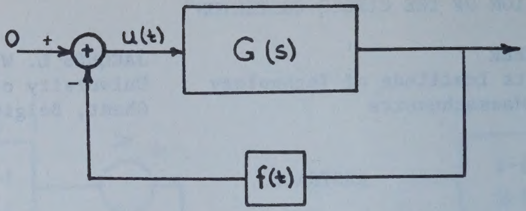


Fig.1: The System under consideration

As is well known the equations of motion for such a sytem can be written as

$$\begin{aligned} \dot{\underline{x}}(t) &= \underline{A}\underline{x}(t) + \underline{b}u(t) \\ y(t) &= \underline{c}'\underline{x}(t) \\ u(t) &= -f(t)y(t) \end{aligned} \tag{3}$$

where $\underline{x}(t)$ is an n-dimensional vector whose successive components are time derivatives of a scalar x(t), i.e. letting D = d/dt

$$\underline{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ Dx(t) \\ \cdot \\ \cdot \\ \cdot \\ D^{n-2}x(t) \\ D^{n-1}x(t) \end{bmatrix} \tag{4}$$

and $\underline{A}$, $\underline{b}$ and $\underline{c}'$ (' denotes transpose) are given by

$$\underline{A} = \begin{bmatrix} 0 & 1 & 0 & \cdot\cdot\cdot\cdot\cdot & 0 & 0 \\ 0 & 0 & 1 & 0 \cdot\cdot\cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot\cdot\cdot\cdot\cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 \cdot\cdot\cdot\cdot & 0 & 1 \\ -p_o & -p_1 & \cdot & \cdot\cdot\cdot\cdot\cdot & -p_{n-2} & -p_{n-1} \end{bmatrix} ; \underline{b} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} ; \underline{c} = \begin{bmatrix} q_o \\ q_1 \\ \cdot \\ \cdot \\ q_{n-2} \\ q_{n-1} \end{bmatrix} \tag{5}$$

Equation (3) together with (4) and (5) will be taken as the state space desciption of the system shown in figure 1 and alternatively will also be referred to as that system.

The system (4) will be said to be asymptotically stable if the null solution of

$$\dot{\underline{x}}(t) = \underline{A}\underline{x}(t) - f(t)\underline{b}\underline{c}'\underline{x}(t) \tag{6}$$

is asymptotically stable in the sense of Liapunov. Notice that in terms of the scalar x(t) defined by (4) the equation of motion for the system can be written as

$$p(D)x(t) + f(t)q(D)x(t) = 0 \tag{7}$$

The null solution of equation (7) will be said to be asymptotically stable if the null solution of the corresponding vector equation (6) is asymptotically stable.

The main tool to be used in establishing the stability criterion derived in this paper is Liapunov stability theory. A time-varying quadratic Liapunov function will be constructed using an integral representation of the type used by Brockett(1). The essential properties of this representation are,

a) the quantity

$$d\eta = p(D)xq(D)x - [(\mathrm{Ev}\ p(D)q(-D))^{-}x]^{2}dt \qquad (8)$$

is an exact differential if Ev p(D)q(-D) denotes the even part of p(D)q(-D) and $(\ )^{-}$ denotes the right half plane spectral factor.

b) the quadratic form

$$V(\underline{x})=\underline{x}'\underline{Q}\underline{x}= \int_{t(\underline{0})}^{t(\underline{x})} p(D)xq(D)x - [(\mathrm{Ev}\ p(D)q(-D))^{-}x]^{2}dt \qquad (9)$$

is positive definite if q(s)/p(s) is positive real and all the zeros of p(s) lie in the half plane Re s < 0.

c) Clearly $\dot{V}(\underline{x})$ is given by

$$\dot{V}(\underline{x}) = 2\dot{\underline{x}}'\underline{Q}\underline{x} = p(D)xq(D)x - [(\mathrm{Ev}\ p(D)q(-D))^{-}x]^{2} \qquad (10)$$

d) If in addition to satisfying the conditions of b) above q(s) and p(s) are such that for the polynomials $u_i(s)$

$$\mathrm{Re}\ [p(j\omega)q(-j\omega) - \sum_{i=1}^{k} u_i(j\omega)u_i(-j\omega)] \geq 0$$

the quadratic form

$$V_1(\underline{x}) = \underline{x}'\underline{K}\underline{x} = \int_{t(\underline{0})}^{t(\underline{x})} p(D)xq(D)x - \sum_{i=1}^{k} (u_i(D)x)^{2} - [(\mathrm{Ev}\ p(D)q(-D) - \sum_{i=1}^{k} u_i(D)u_i(-D))^{-}x]^{2}dt \qquad (11)$$

is positive definite.*

e) $\dot{V}_1(\underline{x})$ is given by

$$\dot{V}_1(\underline{x}) = 2\underline{x}'\underline{K}\underline{x} = p(D)xq(D)x - \sum_{i=1}^{k} (u_i(D)x)^{2} - [(\mathrm{Ev}\ p(D)q(-D) - \sum_{i=1}^{k} u_i(D)u_i(-D))^{-}x]^{2} \qquad (12)$$

Keeping in mind that $\underline{x} = \mathrm{col}\ [x\ ;\ Dx\ ;\ D^{2}x\ ;\ \ldots\ldots\ D^{n-1}x]$ the integrals in equations (9) and (11) are to be thought of as line integrals in the state space along any smooth curve which joins the point $\underline{0}$ with the point $\underline{x}$, parametrized by t.

---

* This result is not proven directly in reference [1]. However it is only an extension of corollary 2 and can be proven in an analogous manner.

Lemma 1: Let $m(D) = D^r + m_{r-1}D^{r-1} + \dots + m_1 D + m_0$. If the null solution of

$$m(D)p(D)w + f(t)m(D)q(D)w = 0 \qquad (13)$$

is asymptotically stable then the null solution of

$$p(D)x + f(t)q(D)x = 0 \qquad (7)$$

is asymptotically stable.

Proof: In terms of the state vector $\underline{w}(t) = \mathrm{col}[w, Dw, \dots D^{r+n-1}w]$ the equation (13) can be written as the vector differential equation

$$\dot{\underline{w}}(t) = \underline{B}\underline{w}(t) - f(t)\underline{b}\underline{c}'\underline{w}(t) \qquad (14)$$

where

$$\underline{B} = \begin{bmatrix} 0 & 1 & 0 & . & . & . & 0 & 0 \\ 0 & 0 & 1 & 0 & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 0 & 1 \\ -g_o & -g_1 & -g_2 & . & . & -g_{n-2} & -g_{n-1} \end{bmatrix} ; \ \underline{b} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \\ 1 \end{bmatrix} ; \ \underline{c} = \begin{bmatrix} h_o \\ h_1 \\ . \\ . \\ . \\ h_{n-2} \\ h_{n-1} \end{bmatrix} \qquad (15)$$

and $g(D) = m(D)p(D) = D^{n+r} + g_{n+r-1}D^{n+r-1} + \dots + g_1 D + g_o$

$h(D) = m(D)q(D) = h_{n+r-1}D^{n+r-1} + h_{n+r-2}D^{n+r-2} + \dots + h_o$

Letting $y(t) = \underline{c}'\underline{w}(t)$ and $u(t) = -f(t)y(t)$ equation (14) can be written as

$$\begin{aligned} \dot{\underline{w}}(t) &= \underline{B}\underline{w}(t) + \underline{b}u(t) \\ y(t) &= \underline{c}'\underline{w}(t) \\ u(t) &= -f(t)y(t) \end{aligned} \qquad (16)$$

From (15) it is clear that the system (16) is controllable. Hence there exists a non-singular transformation such that (see Kalman(2))

$$\underline{P}\underline{w} = \underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix} \qquad (17)$$

$$\begin{aligned} \dot{\underline{z}}(t) &= \underline{E}\underline{z}(t) + \underline{\beta}u(t) \\ y(t) &= \underline{\gamma}'\underline{z}(t) \\ u(t) &= -f(t)y(t) \end{aligned} \qquad (18)$$

where $\underline{z}_1$ represents the controllable but unobservable states, $\underline{z}_2$ represents the controllable and observable states, and

$$\underline{E} = \begin{bmatrix} \underline{E}_{11} & \underline{E}_{12} \\ \underline{0} & \underline{E}_{22} \end{bmatrix} \; ; \; \underline{\beta} = \begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} \; ; \; \underline{\gamma} = \begin{bmatrix} \underline{0} \\ \underline{c}_2 \end{bmatrix} \tag{19}$$

Clearly the transfer function $G(s) = \frac{\hat{y}(s)}{\hat{u}(s)}$ where $\hat{y}(s)$ and $\hat{u}(s)$ are the Laplace transforms of $y(t)$ and $u(t)$ respectively is given by

$$G(s) = \underline{\gamma}'(\underline{I}_{n+r}s - \underline{E})^{-1}\underline{\beta} = \underline{c}_2(\underline{I}_n s - \underline{E}_{22})^{-1}\underline{\beta}_2 = \frac{q(s)}{p(s)} \tag{20}$$

where $\underline{I}_j$ = j by j unit matrix.

Hence it follows that $\underline{z}_2$ is also a state vector for the system of figure 1 and consequently there exists a non-singular transformation $\underline{R}$ such that

$$\underline{x}(t) = \underline{R}\underline{z}_2(t) \tag{21}$$

Since the stability properties of the null solution of equation (16) remain invariant under the transformations (17) and (21) it follows that if the null solution of equation (13) and hence the null solution of equation (14) is asymptotically stable the null solution of equation (4) will also be asymptotically stable. Consequently it follows immediately that the null solution of equation (7) is asymptotically stable.

Q.E.D.

Lemma 2: Let $m(s)$ be given as in lemma 1 and let $\underline{x}(t)$ redefined as

$$\underline{x}(t) = \text{col}\,[x(t);\, Dx(t);\, \ldots\ldots\ldots;\, D^{n+r-1}x(t)]$$

be the state vector correspondig to the equation

$$m(D)p(D)x(t) + f(t)m(D)q(D)x(t) = 0\,;$$

assume $\frac{n(s)}{m(s)}$ is positive real and the zeros of $m(s)$ lie in the half plane $\text{Re}\, s < 0$. Then the line integral

$$V_2(\underline{x}) = \underline{x}'\underline{L}\underline{x} = e^{-2\alpha t(\underline{x})} \int_{t(\underline{0})}^{t(\underline{x})} m(D)[(q(D)x)e^{\alpha t}]n(D)[(q(D)x)e^{\alpha t}]$$

$$- [(\text{Ev}\; m(D)n(-D))^{-}(q(D)x)e^{\alpha t}]^2 \, dt \tag{22}$$

is independent of path and the quadratic form $\underline{x}'\underline{L}\underline{x}$ is a positive semi-definite function of $\underline{x}$.

Proof: Let $\theta = (q(D)x)e^{\alpha t}$ and let the vector $\underline{\theta} = \text{col}[\theta\,; D\theta;\, ..;D^{k-1}\theta]$ where k=r if the degree of m(s) is greater than or equal to the degree of n(s) and k = r+1 if the degree of n(s) exceeds that of m(s). Consider

$$\theta_i = D^{i-1}\,\theta = D^{i-1}[\,(q(D)x)e^{\alpha t}] \tag{23}$$

Clearly

$$\theta_i = [\,(D+\alpha)^{i-1}q(D)x]e^{\alpha t} \tag{24}$$

Hence since the degree of q(D) is less than that of p(D) it follows that

$$\underline{\theta}(t) = \underline{M}\ \underline{x}(t)\ e^{\alpha t} \tag{25}$$

where $\underline{M}$ is a k by n+r matrix

Let $\underline{\theta} = \underline{M}\ \underline{x}\ e^{\alpha t(\underline{x})}$ and consider the quadratic form

$$\underline{\theta}'\underline{K\theta} = \int_{t(\underline{0})}^{t(\underline{\theta})=t(\underline{x})} m(D)\theta n(D)\theta - [(\mathrm{Ev}\ m(D)n(-D))^{-}\theta]^2\ dt \tag{26}$$

From property b) above it is clear that the line integral (26) is independent of path and $\underline{\theta}'\underline{K\theta}$ is positive definite. Then it follows from the definition of θ and $\underline{\theta}$ that

$$V_2(\underline{x}) = e^{-2\alpha t(\underline{x})}\ \underline{\theta}'\underline{K\theta} = e^{-2\alpha t(\underline{x})}e^{2\alpha t(\underline{x})}\underline{x}'\underline{M}'\underline{KMx} \tag{27}$$

$$= \underline{x}'\underline{Lx}$$

where $\underline{L} = \underline{M}'\underline{KM}$ is a positive semidefinite n+r by n+r matrix.
This proves the lemma.

A STABILITY CRITERION FOR TIME-VARYING SYSTEMS

In the statement of the theorem below the following definition will be used,

Definition: A rational function $Z(s) = \frac{n(s)}{m(s)}$ will be called strictly positive real (s.p.r.) if Z(s) is positive real, the zeros of m(s) and n(s) lie all in the half plane Re s< 0, and Re Z(jω) > 0 for all finite ω.

Theorem: The system shown in figure 1 is asymptotically stable if there exists a positive real Z(s) = n(s)/m(s) such that

$$\text{(i)}\quad Z(s)\ (G(s) + \frac{1}{k}\ )\ \text{and}\ Z(s-\alpha)\ \text{are s.p.r.}$$

$$\text{(ii)}\quad 0 \leq f(t) < k$$

$$\text{(iii)}\quad \dot{f}(t) \leq 2\alpha f(t)[1 - \frac{f(t)}{k}]$$

Proof: Since the equation of motion for the system shown in figure 1 is given by (7) in view of lemma 1 it is sufficient to show that under the assumptions of the theorem the null solution of the equation

$$m(D)p(D)x + f(t)m(D)q(D)x = 0 \tag{28}$$

is asymptotically stable. First consider the case where $0 \leq f(t) < \infty$.

For this case $Z(s)G(s) = \frac{n(s)q(s)}{m(s)p(s)}$ is s.p.r. and consequently if the degree of m(s)p(s) is n there exists an $\varepsilon > 0$ such that

$$\mathrm{Re}\ [m(j\omega)p(j\omega)n(-j\omega)q(-j\omega) - \varepsilon(1+\omega^2+\omega^4+\ ....+\ \omega^{2n-2})] \geq 0$$

Consequently the quadratic form

$$V_1(\underline{x}) = \underline{x}'\underline{Kx} = \int_{t(\underline{0})}^{t(\underline{x})} m(D)p(D)xn(D)q(D)x - \varepsilon(x^2+(Dx)^2+ \;....\; +(D^{n-1}x)^2)$$

$$- [(Ev\ m(D)p(D)n(-D)q(-D) - \varepsilon(1-D^2+ \;...\; +(-1)^{n-1}D^{2n-2}))^{-}x]^2\ dt \quad (30)$$

is positive definite, where it will be remembered that

$$\underline{x} = \mathrm{col}[x\ ;\ Dx\ ;\ D^2x\ ;\ .....\ ;\ D^{n-1}x]$$

Also since $\frac{n(s-\alpha)}{m(s-\alpha)}$ is s.p.r. by lemma 2 the quadratic form

$$V_2(\underline{x}) = \underline{x}'\underline{Lx} = e^{-2\alpha t(\underline{x})}\int_{t(\underline{0})}^{t(\underline{x})} n(D-\alpha)((q(D)x)e^{\alpha t})m(D-\alpha)((q(D)x)e^{\alpha t}$$

$$-[(Ev\ m(D-\alpha)n(-D-\alpha)]^{-}(q(D)x)e^{\alpha t}]^2dt \quad (31)$$

will be positive semidefinite.

Consider now the time varying quadratic function

$$V(\underline{x},t) = V_1(\underline{x}) + f(t)V_2(\underline{x}) \quad (32)$$

Clearly $V(\underline{x},t)$ is positive definite, radially unbounded, and in view of the the fact that f(t) is bounded also decrescent. Hence if it can be shown that $\dot{V}(\underline{x},t)$ is negative definite along solutions of equation (28) the asymptotic stability of the null solution of equation (28) will be assured (see Hahn [3]). From equation (32) it follows that along solutions of equation (28)

$$\dot{V}(\underline{x},t) = \dot{V}_1(\underline{x}(t)) + f(t)\dot{V}_2(\underline{x}(t)) + \dot{f}(t)V_2(\underline{x}(t)) \quad (33)$$

Defining

$$r_1(D) \equiv (Evm(D)p(D)n(-D)q(-D) - \varepsilon[1-D^2 +...+(-1)^{n-1}D^{2n-2}])^{-}$$

$$r_2(D) \equiv (Evm(D-\alpha)n(-D-\alpha))^{-}$$

$\dot{V}_1(\underline{x}(t))$ and $\dot{V}_2(\underline{x}(t))$ can be written as

$$\dot{V}_1(\underline{x}(t)) = m(D)p(D)x(t)n(D)q(D)x(t) - \varepsilon[x^2 + (Dx)^2 +...(D^{n-1}x)^2]$$

$$- [r_1(D)x]^2 \quad (34)$$

$$\dot{V}_2(\underline{x}(t)) = e^{-2\alpha t}m(D-\alpha)[(q(D)x)e^{\alpha t}]n(D-\alpha)[(q(D)x)e^{\alpha t}]$$

$$-[r_2(D)[(q(D)x)e^{\alpha t}]]^2$$

$$-2\alpha e^{-2\alpha t}\int_{t(\underline{0})}^{t} m(D-\alpha)[(q(D)x)e^{\alpha\tau}]n(D-\alpha)[(q(D)x)e^{\alpha\tau}] - [r_2(D)[[q(D)x)e^{\alpha\tau}]]^2d\tau -$$

Since $t = t(\underline{x}(t))$ and $m(D)(xe^{\alpha t}) = [m(D+\alpha)x]e^{\alpha t}$

$$\dot{V}_2(\underline{x},t) = m(D)q(D)xn(D)q(D)x - [r_2(D)((q(D)x)e^{\alpha t})]^2 - 2\alpha V_1(\underline{x}(t)) \quad (35)$$

Multiplying equation (28) by $n(D)q(D)x$ it is clear that

$$m(D)p(D)x \quad n(D)q(D)x = -f(t)m(D)q(D)xn(D)q(D)x$$

and consequently $\dot{V}(\underline{x},t)$ along solutions of equation (28) is given by

$$\dot{V}(\underline{x},t) = -\varepsilon[x^2 + (Dx)^2 + \ldots + (D^{n-1}x)^2] - (r_1(D)x)^2$$
$$- f(t)[r_2(D)((q(D)x)e^{\alpha t})]^2 - [2\alpha f(t) - \dot{f}(t)]V_1(\underline{x}(t)) \quad (36)$$

Since $f(t) \geq 0$, $V_1(\underline{x}(t))$ is positive semidefinite and $\varepsilon > 0$ it follows that if

$$\dot{f}(t) \leq 2\alpha f(t) \quad (37)$$

$\dot{V}(\underline{x},t)$ is negative definite thus proving the theorem for $0 \leq f(t) < \infty$. To prove the result for the more general case $0 \leq f(t) < k$ the transformation

$$h(t) = \frac{kf(t)}{k-f(t)} \quad (38)$$

is used. In terms of $h(t)$ the equation of motion for the system then becomes

$$p(D)x + h(t)[q(D) + \frac{1}{k}\, p(D)]x = 0 \quad (39)$$

Since $0 \leq h(t) < \infty$ it follows from the previous result that a sufficient condition for the null solution of equation (39) and hence of equation (28) to be asymptotically stable is that there exists a positive real $Z(s)$ such that

(a) $Z(s)G'(s) = Z(s)[\frac{q(s) + \frac{1}{k}\, p(s)}{p(s)}] = Z(s)\; G(s) + \frac{1}{k}$

and $Z(s - \alpha)$ are s.p.r.

(b) $\dot{h}(t) \leq 2\alpha h(t)$

It is easily verified that condition (b) is equivalent to condition (iii) of the theorem. This completes the proof.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to Prof. R. W. Brockett of the Massachusetts Institute of Technology for helpful suggestions.

REFERENCES

1. Brockett, R. W., "On the Stability of Nonlinear Feedback Systems," IEEE Transaction on Application and Industry, Vol. AP-83, No. 6, pp. 443-448, 1964.

2. Kalman R. E., "Canonical Structure of Linear Dynamical Systems," Proc. of the National Academy of Sciences, Vol. 48, No. 4, pp. 596-600, April, 1962.

3. Hahn, W., Theory and Application of Liapunov's Direct Method, Prentice-Hall, Inc. Englewood Clifs, N.J., 1963.

4. Sandberg, I.W., "A Frequency-Domain for the Stability of Systems Containing a Single Time-Varying Nonlinear Element," Bell Technical Journal, Vol. 43, 1964, pp. 1601-1608.

5. Narendra, K. S. and Goldwyn, R. M., "Generation of Quadratic Type Liapunov Functions for Linear Time-Varying Systems," Technical Report No. 421, Cruft Laboratory, Harvard University, 1963.

6. Bongiorno, J.J. Jr., "Real-Frequency Stability Criteria for Linear Time-Varying Systems," Proceedings of the IEEE, Vol. 52, No. 7, 1964, pp. 832-841.

7. Zames, G., "The Input-Output Stability of Nonlinear and Time-Varying Feedback Systems," Proceedings of the 1964 National Electronics Conference, Chicago, Ill., pp. 725-730.

8. Brockett, R. W. and Forys, L. J., "On the Stability of Systems Containing a Time-Varying Gain," Proceedings of the Second Annual Allerton Conference on Circuit and Systems Theory, University of Illinois, Urbana, Ill., 1964, pp. 413-430.

9. Willems, J. L., "Stability Criteria for Time-Varying Feedback Systems," Unpublished Manuscript.

10. Gruber, M. "The Use of the Shift Theorem to Establish The Stability of Time Varying Systems" Unpublished Manuscript.

PERTURBATION THEORY FOR THE ANALYSIS OF INSTABILITY IN NONLINEAR FEEDBACK SYSTEMS*

JAN C. WILLEMS
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

Although Pliss has demonstrated convincingly that Aizerman's conjecture is not true, his counterexamples do not make clear what the physical and mathematical reasons for the failure of the conjecture are. In this paper a simple and rigorous explanation is given as to why this linearization procedure fails to accurately predict asymptotic stability for nonlinear feedback systems. The results are obtained using the perturbation theory of Cesari and Hale. The existence of periodic solutions is established for a class of systems which on the basis of linearization would appear to be asymptotically stable in the large. The procedure not only establishes counterexamples for Aizerman's "d.c." type of linearization but also for the incremental gain type of linearization.

INTRODUCTION

Consider the feedback system containing a single nonlinearity shown in Fig. 1. The transfer function of the linear forward loop, $G(s)$, is assumed to be the ratio of two polynomials, i.e., $G(s) = \frac{q(s)}{p(s)}$, and the degree of $p(s)$ is assumed to exceed the degree of $q(s)$.

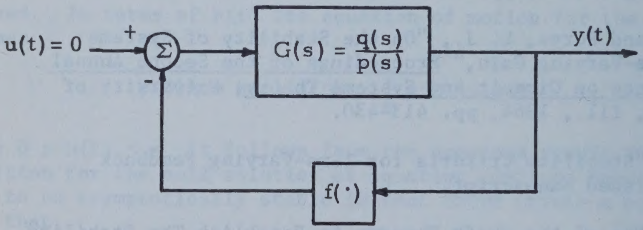


Figure 1: The Feedback System under Consideration

The symbol $f(\cdot)$ indicates an element whose output is $f(\sigma)$ when its input is σ. It is assumed that $f(0) = 0$, and that $f(\sigma)$, together with its first and second partial derivatives with respect to σ, are continuous functions of σ.

If the variable $x(t)$ is related to $y(t)$ by the equation

$$q(D)x(t) = y(t) \; ; \; D = \frac{d}{dt} \tag{1}$$

then the differential equation describing $x(t)$ is:

$$p(D)x(t) + f(q(D)x(t)) = 0 \tag{2}$$

---

[*] This work was supported in part by the National Aeronautics and Space Agency through NASA Research Gaant NGR 22-009-124 with the M.I.T. Electronic Systems Laboratory.

The solution $x(t) \equiv 0$ is called the null-solution of equation (2). The null-solution is said to be asymptotically stable in the large (ASIL) if it is stable (in the sense of Liapunov) and all solutions of (2) converge to the null-solution for $t \to \infty$. Clearly if the null-solution of equation (2) is ASIL, then all $y(t)$ approach zero for $t \to \infty$. Therefore the feedback system shown in Fig. 1 is said to be ASIL if the null-solution of equation (2) is ASIL.

Figure 2 shows the block diagram of what is called "the linearized feedback system associated with the nonlinear feedback system of Fig. 1".

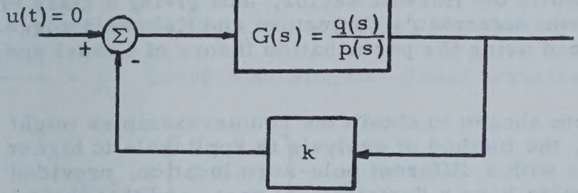


Figure 2: The Linearized Feedback System Associated with the Nonlinear Feedback System Shown in Figure 1

Using root-locus techniques, the Nyquist stability criterion or a Routh-Hurwitz test, the range of k for which this system is asymptotically stable can be determined in a routine fashion.

In engineering applications, the answer to the question as to whether a particular nonlinear feedback system is ASIL is often derived by considering a linearized model. Two common types of linearization are the "d.c." type in which the nonlinearity $f(\sigma)$ is replaced by the "total gain", i.e., $k_\sigma = \frac{f(\sigma)}{\sigma}$, and the "a.c." type in which the nonlinearity is replaced by the "incremental gain", i.e., $k_\sigma = \frac{\partial f(\sigma)}{\partial \sigma}$. The original nonlinear system is then (heuristically) concluded to be ASIL if the linearized model is asymptotically stable for all k_σ.

The gain of a nonlinearity is said to be completely contained in the Hurwitz sector of a nonlinear system with a single nonlinearity if $k_1 < k_\sigma < k_2$ and if the associated linearized system is asymptotically stable for all $k_1 < k < k_2$.

In 1949, the Russian author Aizerman [1] made the conjecture that a feedback system with a single nonlinearity is ASIL if the total gain of the nonlinearity is completely contained in the Hurwitz sector, thus suggesting the "d.c." type of linearization as a test for stability.

It is possible to show that for second order systems the conjecture is true, with the exception, however, of some cases where the nonlinearity approaches a boundary of the Hurwitz sector asymptotically and the corresponding solution is unbounded. The counter-example given by Krasovskii [2] is, in fact, of this kind. In 1958 Pliss [3] gave counterexamples of a much more satisfactory type. But the very stringent conditions on the nonlinearity and the involved mathematics kept the work of Pliss from being well known. More recently, Dewey and Jury [4], and Fitts [5] gave numerical counterexamples using computer simulation.

Kalman [6] has suggested that the nonlinear feedback system under consideration is ASIL if the incremental gain of the nonlinearity is completely contained in the Hurwitz sector. Even though this "a.c." type of linearization predicts asymptotic stability in the large only for a subclass of the nonlinearities for which Aizerman's conjecture does, this suggestion is not true in general either. Fitts [5] gives counterexamples derived from a computer simulation.

It is the purpose of this paper to present a simple, rigorous proof of the existence of oscillations in a fourth order system. It will be shown that the total gain as well as the incremental gain of the nonlinearity can be completely contained in the Hurwitz sector, thus giving a class of counterexamples to both Aizerman's conjecture and Kalman's suggestion. The results are obtained using the perturbation theory of Cesari and Hale [7].

Although the system chosen to obtain the counterexamples might seem to be quite particular, the method of analysis is applicable to higher order systems or to systems with a different pole-zero location, provided that at least two pairs of poles have a "small" real part, and that their imaginary parts are rationally related. These examples show the need for caution in applying linearization techniques in stability analysis.

Much recent research has been devoted to finding sufficient conditions for asymptotic stability in the large for nonlinear systems containing a single nonlinearity. The study of a particular class of systems as done in this paper may suggest to what extent and for which cases the existing criteria can be improved.

PROBLEM STATEMENT

The fourth order system to be considered is a particular case of the class of nonlinear feedback systems shown in Fig. 1. The system has a transfer function given by

$$G(s) = \frac{s^2}{(s^2+1)(s^2+9) + \epsilon(\alpha s^3 + \beta s^2 + \gamma s + \delta)} \tag{3}$$

and a nonlinear feedback $\epsilon f(\cdot)$. Throughout the analysis, $f(\sigma)$ will be assumed to be a continuous function of σ, together with its first and second partial derivatives. The differential equation describing this feedback system is:

$$x^{(4)} + 10x^{(2)} + 9x + \epsilon(\alpha x^{(3)} + \beta x^{(2)} + \gamma x^{(1)} + \delta x) + \epsilon f(x^{(2)}) = 0 \tag{4}$$

with $\alpha, \beta, \gamma, \delta, \epsilon$ constants, and $x^{(i)} = \dfrac{d^i x}{dt^i}$.

Conditions on $f(\cdot)$, $\alpha, \beta, \gamma, \delta$ will be established such that for ϵ sufficiently small equation (4) has a periodic solution other than the null-solution.

In particular it will be shown that the restrictions on $f(\cdot)$ and $\alpha, \beta, \gamma, \delta$ do not exclude the possibility of a non-trivial periodic solution when $k_\sigma = \dfrac{f(\sigma)}{\sigma}$, as well as $k_\sigma = \dfrac{\partial f(\sigma)}{\partial \sigma}$ are such that the linearized system is asymptotically stable for all k_σ.

STATE-SPACE REPRESENTATIONS OF THE SYSTEM

Equation (4) can be expressed as the following system of first order differential equations:

$$\underline{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -9 & 0 & -10 & 0 \end{bmatrix} \underline{x} - \epsilon \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \delta & \gamma & \beta & \alpha \end{bmatrix} \underline{x} - \epsilon \begin{bmatrix} 0 \\ 1 \\ 0 \\ -10 \end{bmatrix} f(x_1) + \underline{O}(\epsilon) \qquad (5)$$

where $\underline{x}$ denotes a column vector $\underline{x} = \text{col}\,(x_1, x_2, x_3, x_4)$, x of equation (4) and $\underline{x}$ are related by $x_1 = \ddot{x}$, and $\underline{O}(\epsilon)$ denotes a vector such that $\lim\limits_{\epsilon \to 0} \frac{\underline{O}(\epsilon)}{\epsilon} = \underline{0}$. Using a nonsingular linear transformation,

$$\underline{z} = \frac{1}{8} \begin{bmatrix} 9 & 0 & 1 & 0 \\ 0 & 9 & 0 & 1 \\ -1 & 0 & -1 & 0 \\ 0 & -\frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix} \underline{x} \qquad (6)$$

The following differential equation in $\underline{z}$ is obtained

$$\dot{\underline{z}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & -3 & 0 \end{bmatrix} \underline{z} + \epsilon \begin{bmatrix} 0 \\ -1 \\ 0 \\ \frac{1}{3} \end{bmatrix} f_1(\underline{z}) + \epsilon \begin{bmatrix} 0 \\ 1 \\ 0 \\ -3 \end{bmatrix} f_2(\underline{z}) + \underline{O}(\epsilon) \qquad (7)$$

where $f_1(\underline{z}) = \frac{1}{8} [(\delta-\beta); (\gamma-\alpha); (\delta-9\beta); 3(\gamma-9\alpha)]\, \underline{z}$ (8a)

$$f_2(\underline{z}) = \frac{1}{8} f(z_1 + z_3) \qquad (8b)$$

Equation (7) shows that the system represents two harmonic oscillators with frequencies $1/2\pi$ and $3/2\pi$, coupled by terms of order ϵ. Consequently, one expects that a periodic solution will contain a component of frequency $\frac{1}{2\pi}(1+\mu(\epsilon))$, a component of frequency $\frac{3}{2\pi}(1+\mu(\epsilon))$, and higher harmonics, with $\mu(\epsilon)$ and the amplitude of the higher harmonics of order ϵ.

In order to obtain a differential equation for which averaging theory can be used, the system should be brought into the form:

$$\dot{\underline{x}} = \underline{X}(\underline{x}, t, \epsilon) \; ; \; \underline{X}(\underline{x}, t, 0) = \underline{0}$$

There are different ways of doing this (see Hale [7]). A convenient choice is to define the new variables θ, r_1, r_2, ϕ as a function of $\underline{z}$ as follows:

$$\underline{z} = \begin{bmatrix} r_1 & \sin\theta \\ r_1 & \cos\theta \\ r_3 & \sin(3\theta+\phi) \\ r_3 & \cos(3\theta+\phi) \end{bmatrix} \tag{9}$$

Upon substitution in (5), and after some manipulations, the following system of first order differential equations in θ, r_1, r_3, and ϕ is obtained:

$$\begin{aligned}
\dot{\theta} &= 1 + \epsilon r_1^{-1} \sin\theta\,[f_1(\theta, r_1, r_3, \phi) - f_2(\theta, r_1, r_3, \phi)] + O(\epsilon) \\
\dot{r}_1 &= -\epsilon \cos\theta\,[f_1(\theta, r_1, r_3, \phi) - f_2(\theta, r_1, r_3, \phi)] + O(\epsilon) \\
\dot{r}_3 &= \epsilon \cos(3\theta+\phi) \left[\frac{f_1(\theta, r_1, r_3, \phi)}{3} - 3f_2(\theta, r_1, r_3, \phi)\right] + O(\epsilon) \qquad (10) \\
\dot{\phi} &= -\epsilon\{3r_1^{-1} \sin\theta\,[f_1(\theta, r_1, r_3, \phi) - f_2(\theta, r_1, r_3, \phi)] \\
&\quad + r_3^{-1} \sin(3\theta+\phi) \left[\frac{f_1(\theta, r_1, r_3, \phi)}{3} - 3f_2(\theta, r_1, r_3, \phi)\right]\} + O(\epsilon)
\end{aligned}$$

where $f_1(\theta, r_1, r_3, \phi)$ and $f_2(\theta, r_1, r_3, \phi)$ are given by (8) and (9).

For ϵ sufficiently small and r_1, r_3 bounded, t can be eliminated from these equations to obtain

$$\frac{dr_1}{d\theta} = \epsilon \frac{-\cos\theta(f_1 - f_2)}{1+\epsilon r_1^{-1} \sin\theta(f_1 - f_2)} + O(\epsilon) \tag{11a}$$

$$\frac{dr_3}{d\theta} = \epsilon \frac{\cos(3\theta+\phi)(\frac{f_1}{3} - 3f_2}{1+\epsilon r_1^{-1} \sin\theta(f_1 - f_2)} \tag{11b}$$

$$\frac{d\phi}{d\theta} = \epsilon \frac{-\{3r_1^{-1} \sin\theta(f_2 - f_1) + r_3^{-1} \sin(3\theta+\phi)(\frac{f_1}{3} - 3f_2)\}}{1 + \epsilon r_1^{-1} \sin\theta(f_1 - f_2)} + O(\epsilon) \tag{11c}$$

THE DETERMINING EQUATIONS

The following theorem and corollary are basic results from averaging theory, and are presented here for easy reference.

Theorem 1: Given a system of differential equations:

$$\frac{d\underline{z}}{d\theta} = \epsilon\, \underline{Z}(\theta, \underline{z}, \epsilon) \tag{12}$$

where $\underline{Z}(\theta+T, \underline{z}, \epsilon) = \underline{Z}(\theta, \underline{z}, \epsilon)$

and $\underline{Z}(\theta, \underline{z}, \epsilon)$ has continuous first and second partial derivatives with respect to $\theta, \underline{z}$, and ϵ for $-\infty < \theta < \infty$, $0 < \|\underline{z}\| < R$, $0 < |\epsilon| < \epsilon_o$ with $R, \epsilon_o > 0$

and given the vector $\underline{Z}_o(\underline{z}, \epsilon)$ defined as:

$$\underline{Z}_o(\underline{z}, \epsilon) = \frac{1}{T} \int_o^T \underline{Z}(\theta, \underline{z}, \epsilon)\, d\theta \tag{13}$$

if there exists an $\underline{a}(\epsilon)$, $\|\underline{a}(\epsilon)\| < R$, for $0 \leq |\epsilon| \leq \epsilon_2$, such that

$$\underline{Z}_o(\underline{a}(\epsilon), \epsilon) = \underline{0} \tag{14}$$

then there exists an $\epsilon_1 > 0$ such that (12) has a periodic solution $\underline{z}(t, \epsilon), \underline{z}(t+T, \epsilon) = \underline{z}(t, \epsilon)$, for all $0 \leq |\epsilon| \leq \epsilon_1$. Moreover, this periodic solution is a continuous function of ϵ.

Proof: For a proof of this theorem, see e.g., the book by Hale [7].

Corollary: If there exists a vector $\underline{a}_o$, $\|\underline{a}_o\| < R$ such that

$$\underline{Z}_o(\underline{a}_o, 0) = \underline{0} \tag{15}$$

$$\det \left[\frac{\partial \underline{Z}_o(\underline{z}, 0)}{\partial \underline{z}} \right]_{\underline{z} = \underline{a}_o} \neq 0$$

then there exists an $\epsilon_1 > 0$ such that (12) has a periodic solution, $\underline{z}(\theta, \epsilon), \underline{z}(\theta+T, \epsilon) = \underline{z}(\theta, \epsilon)$, for $0 \leq |\epsilon| \leq \epsilon_1$. Furthermore $\lim_{\epsilon \to o} \underline{z}(\theta, \epsilon) = \underline{a}_o$.

The equations represented by equation (15) are called the determining equations.

Proof: This corollary follows from the preceding theorem and the implicit-function theorem [7].

System (11) is a special case of (12) since the right-hand side of (11) is periodic in θ with period 2π, and the continuity conditions are satisfied. The determining equations for this system become:

$$(\gamma-a)R_1' - \frac{1}{\pi}\int_0^{2\pi} f(R_1'\sin\theta + R_3'\sin(3\theta+\Phi'))\cos\tau\, d\tau = 0 \qquad (16a)$$

$$\frac{(\gamma-9a)}{3}R_3' - \frac{1}{\pi}\int_0^{2\pi} f(R_1'\sin\theta + R_3'\sin(3\theta+\Phi'))\cos(3\theta+\Phi')\, d\theta = 0 \qquad (16b)$$

$$(\frac{10\delta}{3} - 2\beta)R_1'R_3' - \frac{1}{\pi}\int_0^{2\pi} f(R_1'\sin\theta+R_3'\sin(3\theta+\Phi'))(R_3'\sin\theta+R_1'\sin(3\theta+\Phi'))d\theta = 0 \qquad (16c)$$

This system of algebraic equations in R_1', R_3', and Φ' will symbolically be represented by $\underline{F}(R_1', R_3', \Phi') = \underline{0}$.

Theorem 2: If the system of algebraic equations $\underline{F}(R_1', R_3', \Phi') = \underline{0}$ has a solution R_1, R_3, Φ with $R_1R_3 \neq 0$, and if

$$\det\left[\frac{\partial\underline{F}(R_1', R_3', \Phi')}{\partial(R_1', R_3', \Phi')}\right]_{R_1' = R_1, R_3' = R_3, \Phi' = \Phi} \neq 0 \qquad (17)$$

then there exists an $\epsilon_1 > 0$ such that the system of differential equations (11) has a periodic solution (other than the null-solution), with period 2π, for $0 \leq |\epsilon| \leq \epsilon_1$. Moreover, the limit for $\epsilon \to 0$ of this periodic solution is R_1, R_3, Φ.

Proof: This is a direct application of the corollary.

Remarks: Equation (16a) expresses the fact that the phase shift on the fundamental component of the periodic solution through the linear and nonlinear elements must be matched and equation (16b) expresses this fact for the third harmonic.

The only thing that is left to be proven is that this periodic solution of system (11) yields a periodic solution of the original differential equation.

Theorem 3: If the conditions of the preceding theorem are satisfied, then there exists an $\epsilon_1' > 0$ such that the differential equation (4) has a non-trivial periodic solution for $0 \leq |\epsilon| \leq \epsilon_1'$.

Proof: Upon substitution of the periodic solution of system (11) into the first equation of (10), the following differential equation relating θ and t is obtained

$$\frac{d\theta}{dt} = 1 + \epsilon r_1^{-1}(\theta,\epsilon)\sin\theta\,[f_1(\theta, r_1(\theta,\epsilon), r_3(\theta,\epsilon), \phi, \epsilon)) - f_2(\theta, r_1(\theta,\epsilon), r_3(\theta,\epsilon), \phi(\theta,\epsilon))] + O(\epsilon) \qquad (18)$$

since $r_1(\theta,\epsilon), r_3(\theta,\epsilon), \phi(\theta,\epsilon)$ are periodic in θ with period 2π, this equation can be written as

$$\frac{d\theta}{dt} = 1 + \epsilon g(\theta,\epsilon) \tag{19}$$

with $g(\theta+2\pi,\epsilon) = g(\theta,\epsilon)$

Defining $$g_o(\epsilon) = \frac{1}{2\pi}\int_0^{2\pi} g(\theta,\epsilon)\,d\theta \tag{20}$$

Equation (19) can be written as

$$\frac{d[\theta-(1+\epsilon g_o(\epsilon))t]}{dt} = \epsilon\,(g(\theta,\epsilon)-g_o(\epsilon)) \tag{21}$$

and letting $\theta' = \theta - (1+\epsilon g_o(\epsilon))t$ equation (21) becomes

$$\frac{d\theta'}{dt} = \epsilon[g(\theta'+(1+\epsilon g_o(\epsilon))t) - g_o(\epsilon)] \tag{22}$$

The right hand side of (22) is periodic in t with period $T' = 2\pi/(1+\epsilon g_o(\epsilon))$, and it can be verified that the continuity conditions are satisfied. In this particular case, the function corresponding to $\underline{a}(\epsilon)$ of Theorem 1 is an arbitrary continuous function of ϵ. The right hand side of (22) has a zero mean and from Theorem 1 it is thus concluded that there exists an $\epsilon_1' > 0$, such that the solution of (22), $\theta'(t,\epsilon)$, is periodic, $\theta'(t+T',\epsilon) = \theta'(t,\epsilon)$, for $0 \leq |\epsilon| \leq \epsilon_1'$. Thus it follows that

$$\theta = t + \epsilon g_o(\epsilon)t + \theta'(t,\epsilon) \tag{23}$$

Considering (9) and (23) leads to the conclusion that system (7), and thus the original differential equation (4), has a non-trivial periodic solution for $0 \leq |\epsilon| \leq \epsilon_1'$. The period of this solution is T', and as a function of R_1, R_3, Φ is given by

$$T' = 2\pi\left[1 + \epsilon\,\frac{1}{16R_1}(\beta-\delta) + \frac{1}{\pi}\int_0^{2\pi} f(R_1\sin\theta + R_3\sin(3\theta+\omega))\sin\theta\,d\theta + O(\epsilon)\right] \tag{24}$$

To summarize, it has been shown that if the algebraic equations (16) have a solution R_1, R_3, Φ with $R_1 R_3 \neq 0$, and if (17) is satisfied, then there exists an $\epsilon_1' > 0$ such that the nonlinear differential equation (4) has a periodic solution other than the null-solution for $0 \leq |\epsilon| \leq \epsilon_1'$. Moreover, the limit for $\epsilon \to 0$ of this periodic solution is

$$x(t) = -R_1 \sin t - \frac{R_3}{9}\sin(3t+\Phi) \tag{25}$$

and the period is given by (24).

The poles of the linearized system associated with nonlinear system (3) are given by the zeros of the polynomial

$$s^4 + 10s^2 + 9 + \epsilon(\alpha s^3 + \beta s^2 + \gamma s + \delta) + \epsilon k_\sigma s^2 \tag{26}$$

where k_σ is given by $\frac{f(\sigma)}{\sigma}$ if the "d.c." type of linearization is used and by $\frac{\partial f(\sigma)}{\partial \sigma}$ for the "a.c." type of linearization. It is easily verified that for ϵ sufficiently small, and the input to the nonlinearity bounded (and thus k_σ bounded), the following zero-locations of polynomial (26) occur:

(i) if $\epsilon > 0 \quad \alpha > 0 \quad \gamma > 0 \quad (\gamma-\alpha)(\gamma-9\alpha) < 0$ (27a)

or $\epsilon > 0 \quad \alpha < 0 \quad \gamma > 0 \quad (\gamma-\alpha)(\gamma-9\alpha) < 0$ (27b)

then the zeros of (26) are in $\mathrm{Re}\, s < 0$

(ii) if $\epsilon < 0 \quad \alpha > 0 \quad \gamma > 0 \quad (\gamma-\alpha)(\gamma-9\alpha) < 0$ (28a)

or $\epsilon > 0 \quad \alpha < 0 \quad \gamma < 0 \quad (\gamma-\alpha)(\gamma-9\alpha) < 0$ (28b)

then the zeros of (26) are in $\mathrm{Re}\, s > 0$

(iii) if $(\gamma-\alpha)(\gamma - 9\alpha) > 0$ then two zeros of (26) are in $\mathrm{Re}\, s < 0$ and the other two are in $\mathrm{Re}\, s > 0$

Theorem 4: If $f(\cdot), \alpha, \beta, \gamma, \delta$ are such that equations (16) have a solution R_1, R_3, Φ with $R_1 R_3 \neq 0$ and if (17) is satisfied, then there exists an $\epsilon_o > 0$ such that for either $0 < \epsilon < \epsilon_o$ or $-\epsilon_o < \epsilon < 0$

(i) equation (4) has a periodic solution other than the null-solution

(ii) the total gain and the incremental gain of the nonlinearity are completely contained in the Hurwitz sector of equation (4) for all inputs to the nonlinearity.

Proof: Since (16) and (17) are satisfied it follows from Theorem (3) that there exists an $\epsilon_1' > 0$ such that for all $0 \leq |\epsilon| \leq \epsilon_1'$ equation (4) has a periodic solution. This solution is bounded and other than the null-solution since $R_1 R_3 \neq 0$ and since the limit for $\epsilon \to 0$ of this solution is R_1, R_3, Φ. Take $\epsilon_o (\epsilon_o \leq \epsilon_1')$ such that for all inputs to the nonlinearity due to this periodic solution, the location of the zeros of polynomial (26) is given above. Since (16a) and (16b) are satisfied, it follows that

$$(\gamma-\alpha) R_1^2 + (\gamma-9\alpha) R_3^2 = 0 \tag{29}$$

and thus $(\gamma-\alpha)(\gamma-9\alpha) < 0$, and the poles are either in $\mathrm{Re}\, s < 0$ or in $\mathrm{Re}\, s > 0$. Thus the remarkable feature of these periodic solutions is that (for ϵ sufficiently small), they only occur when the linearized system has all its poles either always in the left half plane or always in the right half plane, contrary to what is expected from linearization. Indeed when all poles are always in the LHP, one expects the system to be ASIL, and when they are always in the RHP, one expects all (nontrivial) solutions to be unbounded. Clearly by choosing ϵ of a particular sign, this periodic solution can be made to be a counterexample to both Aizerman's conjecture and Kalman's suggestion. This follows from equation (27). The choice of ϵ is graphically shown in Figure 3.

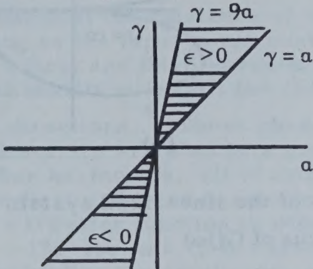


Figure 3: Conditions on ϵ, α, γ to obtain counterexamples to Aizerman's conjecture and Kalman's suggestion

It is interesting to notice (see equation (29)) that in this case the ratio of the amplitudes of the first and third harmonic is (up to order ϵ) independent of the nonlinearity and completely determined by the linear part.

Another remarkable fact is that it follows that with the procedure used here a counterexample to Aizerman's conjecture and Kalman's suggestion can be found for almost any sufficiently smooth nonlinearity. Indeed the only restrictions are that $\alpha, \beta, \gamma, \delta, R_1, R_3, \Phi$ can be chosen such that (16) and (17) are satisfied with $R_1 R_3 \neq 0$. ($f(\cdot)$ linear is the only obvious case where this cannot be done.)

EXAMPLE

Take the particular case where

$$G(s) = \frac{s^2}{(s^2+1)(s^2+9)+\epsilon(\frac{3}{16}s^3 + \frac{1}{4}s^2 + \frac{15}{16}s + \frac{3}{2})} \qquad (30)$$

$$f(\cdot) = (\cdot)^3 \qquad \epsilon > 0 \qquad (31)$$

The differential equation describing this feedback system is

$$x^{(4)} + 10x^{(2)} + 9x + \epsilon(\frac{3}{16}x^{(3)} + \frac{1}{4}x^{(2)} + \frac{15}{16}x^{(1)} + \frac{3}{2}x) + \epsilon(x^{(2)})^3 = 0 \qquad (32)$$

The linearized feedback system associated with the non-linear feedback system given by (30) and (31) is asymptotically stable for all positive feedback. The root-locus for $k \geq 0$ and the Nyquist locus of $G(j\omega)$ are shown in Figure 4.

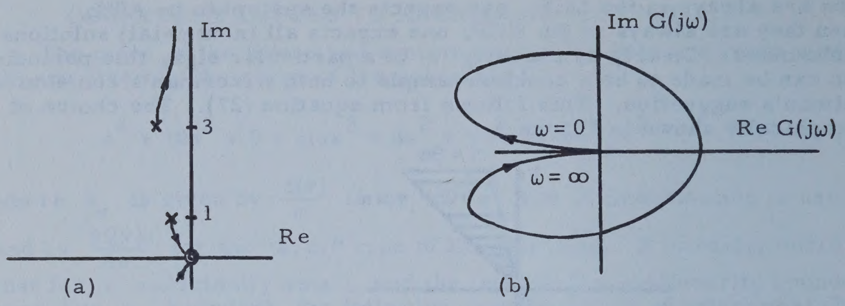


Figure 4: (a) Root-locus of the linearized system for **k** ≥ 0
(b) Nyquist locus of $G(j\omega)$

The total gain, $k_\sigma = \epsilon\sigma^2$, as well as the incremental gain, $k_\sigma = 3\epsilon\sigma^2$ are always positive. Thus both these linearization procedures predict that the null-solution of this equation is ASIL. The determining equations for this case become

$$R_1' + R_1'^2 R_3' \sin \Phi' = 0 \tag{33a}$$

$$R_3' + R_1'^3 \sin \Phi' = 0 \tag{33b}$$

$$18R_1'R_3' - 9R_1'R_3'(R_1'^2 + R_3'^2) + R_1'^2 (R_1'^2 + 3R_3'^2) \cos \Phi' = 0 \tag{33c}$$

This system of equations has a solution $R_1' = 1$, $R_3' = 1$, $\Phi' = -\frac{\pi}{2}$, and it can be verified that the corresponding Jacobian (17) is of full rank. Thus perturbation theory allows to conclude that there exist an $\epsilon_o > 0$ such that for $0 \leq \epsilon \leq \epsilon_0$, equation (32) has a periodic solution, $x(t, \epsilon)$, with period $T(\epsilon)$. Moreover,

$$\lim_{\epsilon \to o} x(t, \epsilon) = -\sin t + \frac{1}{9} \cos 3t \tag{34}$$

$$T(\epsilon) = 2\pi(1 + \frac{13}{64}\epsilon + O(\epsilon)) \tag{35}$$

Clearly this solution violates both Aizerman's conjecture and Kalman's suggestion.

CONCLUSIONS

It was shown that if a solution to the determining equations (16) can be found with $R_1R_3 \neq 0$, and if the Jacobian (17) is of full rank, then the fourth order system described by (3) and (4) has a nontrivial periodic solution for ϵ sufficiently small. This solution is, for ϵ sufficiently small and of a particular sign, a counterexample to both Aizerman's conjecture and Kalman's suggestion. It was also shown that this procedure provides a counterexample for almost any sufficiently smooth nonlinearity.

Although the way the determining equations were derived might seem lengthy, this was essentially only necessary for the sake of rigor. Indeed, these equations are precisely what would be obtained by applying the dual input describing function technique (8). They can, keeping this in mind, be written down almost by inspection.

A physical explanation of the existence of these oscillations is obtained by considering as the input to the nonlinearity a periodic waveform containing a first and third harmonic, and higher harmonics of order ϵ. The nonlinearity will shift the phases of the first and third harmonic in opposite directions.* These phase shifts are of order ϵ. The output of the nonlinearity will also be a periodic waveform, with a first, a third and higher harmonics, all of order ϵ. The linear element shift the phases of the first and third harmonic also in opposite direction, which is a feature of a transfer function in which the phase shift changes from values close to -180° for $\omega = 1$, to values close to $+180^\circ$ for $\omega = 3$. The linear element multiplies the amplitudes of the first and third harmonics by a factor of order ϵ^{-1}, and the higher harmonics by a factor of order 1, thus making it possible to close the loop. An example for $f(\cdot)$ a monotone first and third quadrant function, and a $G(s)$ such that the linearized system is stable for all positive feedback is shown in Figure 5.

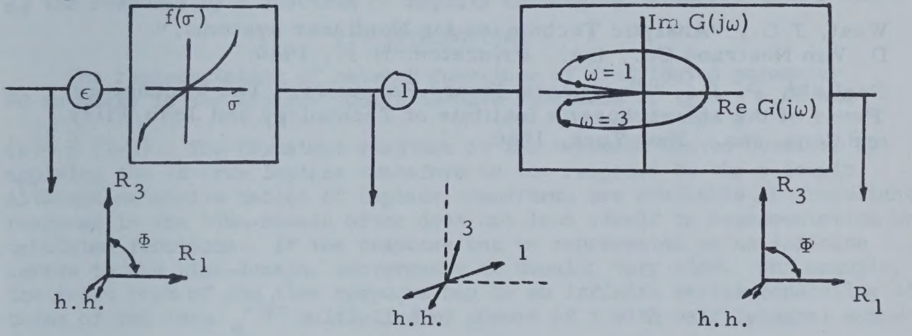


Figure 5: Physical interpretation of the oscillations

ACKNOWLEDGEMENT

The author wishes to thank Prof. R.W. Brockett for his suggestions.

---

* This fact essentially follows from the Manley-Rowe equations (see e.g. [9])- and is expressed here by equation (29).

REFERENCES

1. Aizerman, M.A., "On a Problem Concerning the Stability in the Large of Dynamical Systems," Uspekhi Mat., Nauk 4, 1949, p. 187.
2. Krasovskii, N.N., "Theorems on the Stability of Motions Determined by a System of Two Equations," Prikl. Mat., Mekh. 16, 1953, p. 339.
3. Pliss, V.A., Certain Problems in the Theory of the Stability of Motion, Leningrad University Press, 1958.
4. Dewey, A.G., and Jury, E.T., "A Note on Aizerman's Conjecture," IEEE Trans. on Automatic Control, Vol. AC-10-4, Oct. 1965, p. 482.
5. Fitts, R.E., "Two Counterexamples to Aizerman's Conjecture," IEEE Trans. on Automatic Control, Vol. AC-11-3, July 1966, (to appear).
6. Kalman, R.E., "Physical and Mathematical Mechanisms of Instability in Nonlinear Automatic Control Systems," Trans. ASME, Vol. 79-3, 1957, p. 555.
7. Hale, J.K., "Oscillations in Nonlinear Systems," McGraw-Hill, New York, 1963.
8. West, J.C., "Analytic Techniques for Nonlinear Systems," D. Van Nostrand Co., Inc., Princeton, N.J., 1960.
9. Penfield, P. Jr., "Frequency-Power Formulas," The Technology Press of the Massachusetts Institute of Technology and John Wiley and Sons, Inc., New York, 1960.

RECIPROCAL TIME DOMAIN ANALYSIS OF DISTRIBUTED NETWORKS

W. W. HAPP
NASA
Electronics Research Center
Cambridge, Massachusetts

SOMESHWAR C. GUPTA
Electrical Engineering
Arizona State University
Tempe, Arizona

ABSTRACT

The impulse response in distributed parameter systems results in a spectrum of delta functions in the time domain for $\overline{LC}$ structures. It is shown that for $\overline{RC}$ structures the impulse response results in a spectrum of Poisson derived functions in the reciprocal time domain. This isomorphism is exploited for the analysis of $\overline{RC}$ circuits establishing a calculus of Poisson derived functions. Applications of this technique to thin film networks and minority carrier devices are examined. Illustrative examples of excitations include the impulse, the step and the ramp.

Significant advantages accrue from a spectral description of $\overline{RC}$ networks in the reciprocal time domain: (1) conceptual, by establishing a physical interpretation based on diffusion phenomena (2) analytical, by developing a systematic approach to transients paralleling the network function approach in the frequency domain (3) computational, by expressing the response by a spectrum of rapidly converging terms.

1. INTRODUCTION

The representation of network functions of distributed parameter $\overline{RC}$ networks in the Laplace domain involve functions of type $e^{\pm a\sqrt{s}}$ and $(s)^{\pm\frac{n}{2}}$ [1-3]. The transient response in the t-domain can be obtained by applying the inverse Laplace transform to the response in the s-domain. Although extensive tables of Laplace transforms are available, the resultant response in the time-domain often does not lend itself to representation by tabulated functions. If the response can be represented as an infinite series in the time-domain, convergence is usually very slow. For example, the exact form of the time response may be an infinite series consisting of terms of the form $e^{-B/t}$ multiplied by powers of t with half-integral exponents [2,4,5]; computationally and conceptually difficulties arise: Computationally, the difficulty of manipulating this type of infinite series arises because of lack of criteria for convergence. Conceptually, an interpretation on the basis of a physical model appears futile. For these reasons, the transient response has been estimated by a number of approximation techniques which can be grouped as follows:

(i) Extreme Values of Asymptotic Techniques: The response is computed at intervals very large or very small with respect to the dominant poles [6]. One of these techniques utilizes the generalized initial and final value theorems of the Laplace transform [7].

(ii) Dominant Pole or "Lumping" Techniques: The response is approximated by using an equivalent lumped network which yields a finite number of dominant poles [8]. This technique reduces the problem to lumped network methods and yields closed form solutions.

(iii) Numerical approaches: The expansion of transcendental network functions to yield polynomials results in time response functions which are often accurate only over a limited range and further offer little insight into phenomena governing transients in $\overline{RC}$ circuits.

One promising way of obtaining a numerical solution with adequate accuracy forms the basis of the approach here proposed: An infinite series representation of the time response can be obtained if the impulse response is known. For example a complete and non-redundant set of all network functions of an RC network can be evaluated with the aid of the digital computer. If these functions are plotted, all other responses can be obtained by graphical convolution. A tabulated set of impulse functions will thus be a description of the network in the time-domain with the same generality as a complete matrix of Bode plots describes all possible frequency responses of the network. A major limitation of this approach is its restriction to numerical and graphical representations which are required for each network in every configuration.

A systematic representation of transient response functions based on an analytical rather than numerical or graphical description is the objective of this investigation. To achieve this objective, it was necessary to correlate a prescribed expansion in the s-domain to Poisson distributions in the reciprocal time domain. The utilization of the reciprocal time domain can be compared with the utilization of reciprocal wave lengths and Miller indices in the propagation of waves in periodic structures [9] or the use of the reciprocal momentum space in quantum mechanics [10]. The representation in the reciprocal time domain is facilitated by introducing a new set of functions referred to as G-functions which are closely related to the functions described by Poisson distributions, and the Incomplete Gamma function. The G-functions in reciprocal time domain for $\overline{RC}$ networks play a similar role as the Fourier components for the lumped parameter networks in the frequency domain. It is possible to adapt the techniques here proposed to distributed parameter systems in other branches of science and engineering.

2. DEFINITIONS

For a wide range of distributive parameter networks, the network function N(s), in the Laplace domain can be expressed in the form

$$N(s) = \sum_{n=0}^{\infty} N_n(s) \tag{1}$$

Let

$$\mathcal{L}^{-1}\left\{N_n(s)\right\} = N_n(r) \tag{2}$$

where $\mathcal{L}^{-1}$ represents the inverse Laplace transform and $r = 1/t$. The overall impulse function in the r-domain is, then,

$$N(r) = \sum_{n=0}^{\infty} N_n(r) \tag{3}$$

where $N_n(r)$ can be thought of as spectral component of the impulse response in the reciprocal time domain.

Some simple relationships in s and r domain can be obtained as follows:

1. $\mathcal{L}^{-1}\left\{N_n(as)\right\} = \frac{1}{a} N_n(ar)$, where a is a constant. (4)

2. $\mathcal{L}^{-1}\left\{d^{\mu}/ds^{\mu}\left[N_n(s)\right]\right\} = (-1)^{\mu} \cdot 1/r^{\mu} \cdot N_n(r)$ (5)

3. $\mathcal{L}^{-1}\left\{s^{\mu} N_n(s)\right\} = d^{\mu}/d(\frac{1}{r})^{\mu} \left\{N_n(r)\right\}$ if $N_n^{(k)}(\infty) = 0$, $\mu \geq 0$ for $k = 0,1,2 \ldots \mu-1$. (6)

4. $\mathcal{L}^{-1}\left\{N_n(s)/s^{\mu}\right\} = \int_0^{1/r} \cdots \int_0^{1/r} N_n(r)\, d(1/r)$ if $N_n^{(-k)}(\infty) = 0$, for $k = 1,2,3 \ldots \mu$. (7)

5. $\mathcal{L}^{-1}\left\{N_n(s+a)\right\} = \exp(-a/r)\, N_n(r)$ (8)

The above listed relations are given because of their use in the analysis of distributed parameter networks.

3. DEVELOPMENT OF G-FUNCTION

Definition

The Poisson distribution is defined by

$$p(\lambda,y) = e^{-\lambda}\lambda^{y} / \Gamma(y+1) \tag{9}$$

where $\Gamma x = (x - 1)!$ is the gamma function. Tabulated values are available

Substituting $\lambda = (n+m)^2 r$ and $y = k - 1/2$
into equation (9) yields

$$p\left\{(n+m)^2 r,\ k-1/2\right\} = e^{-(n+m)^2 r} r^{k-1/2} (n+m)^{2k-1} / \Gamma(k+1/2) \tag{10}$$

Now define $G_n(r;m,k) = e^{-(n+m)^2 r} r^{k-1/2} / \Gamma(1/2)$ (11)

or $$G_n(r;m,k) = \rho_k / (n+m)^{2k-1} \cdot p\left\{(n+m)^2 r,\ k-1/2\right\} \tag{12}$$

where $$\rho_k = \Gamma(k+1/2) / \Gamma(1/2) = 2k!\ /\ k!\ 2^{2k} \tag{13}$$

The G-function is a form of the hypergeometric series and, as expected has similar properties to the Poisson distribution.

Properties of G-function

The G-function is defined by four variables n,r,m and k. Certain useful properties are now tabulated.

1. Initial and Final Values

(i) $G_0(r;0,k) = \pi^{-1/2}\ r^{k-1/2}$ (14)

(ii) $G_n(\infty;m,k) = 0$, where $k \leq 0$ is an integer. (15)

(iii) $G_n(0;m,k) = 0$, where $k > 0$ is an integer. (16)

2. Multiplication by r^{α}

$$r^{\alpha} G_n(r;m,k) = G_n(r;m,k+\alpha) \text{ for all } \alpha \tag{17}$$

3. Derivative with respect to 1/r

$$d/d(\tfrac{1}{r})\ \left\{G_n(r;m,k)\right\} = d/dt\left\{G_n(r;m,k)\right\}$$
$$= (n+m)^2 G_n(r;m,k+2) - (k-1/2) G_n(r;m,k+1) \tag{18}$$

Higher order derivatives can be evaluated using recursive relations similar to equation (17).

4. Integration with respect to t where t = 1/r

$$\int_0^t G_n(r;m,k)dt = \sqrt{\pi} \int_0^t e^{-(n+m)^2 r} r^{k-1/2}\ dt \quad \text{where } t = 1/r$$

$$= \sqrt{\pi} \int_r^{\infty} e^{-(n+m)^2 r} r^{k-5/2}\ dr$$

$$= \sqrt{\pi}/(n+m)^{2k-3} \int_{(n+m)^2 r}^{\infty} e^{-q} q^{k-5/2} dq$$

$$= \sqrt{\pi}/(n+m)^{2k-3} \cdot \Gamma(k-3/2,\ (n+m)^2 r) \tag{19}$$

where $\Gamma(a,x) = \int_x^{\infty} e^{-z} z^{a-1} dz$ is an Incomplete Gamma function.

4. BASIC RELATIONSHIPS

From a set of frequently recuring s and r-domain relationships listed in Table 1 as A_n, B_n, C_n and D_n, network functions for the distributed parameter networks [11-14] can be derived and are summarized in Table 2. From the results of Section 2 and using the notation $A_n(r) = \mathcal{L}^{-1}\{A_n(s)\}\Big|_{t=1/r}$ the basic building blocks are obtained:

First $\quad A_n(s) = e^{-2(n+m)\sqrt{s}}$

yields $\quad A_n(r) = (n+m) r^{3/2}\, e^{-(n+m)^2 r} / \Gamma(1/2) = (n+m)\, G_n(r;m,2) \qquad (20)$

TABLE 1

RELATIONSHIPS FOR DISTRIBUTED PARAMETER NETWORKS

	s-domain	r-domain
A_n	$e^{-2(n+m)\sqrt{s}}$	$\Sigma_n (n+m) G_n(r;m,2)$
B_n	$e^{-2(n+m)\sqrt{s}} / \sqrt{s}$	$\Sigma_n G_n(r;m,1)$
C_n	$\sqrt{s}\, e^{-2(n+m)\sqrt{s}}$	$\Sigma_n [(n+m)^2 G_n(r;m,3) - 1/2 G_n(r;m,2)]$
D_n	$s^{-(\gamma+1/2)}$, $\gamma > -1/2$ integer	$\Sigma_n\ 1/\rho_\gamma G_o(r;0;3-2\gamma)$

TABLE 2

s-DOMAIN AND r-DOMAIN RELATIONSHIPS FOR $\overline{\text{RC}}$ FUNCTIONS

L-transform

positive terms	alternating terms	absolute value of terms (sum Σ_n)
$\text{csch} \sqrt{s}$	$\text{sech} \sqrt{s}$	$(2n+1)\, G_n(r;1/2,2)$
$\coth \sqrt{s}$	$\tanh \sqrt{s}$	$n\, G_n(r;0,2) - (n+1)\, G_n(r;1,2)$
$s^{-1/2} \text{csch} \sqrt{s}$	$s^{-1/2} \text{sech} \sqrt{s}$	$2\, G_n(r;1/2,1)$
$s^{-1/2} \coth \sqrt{s}$	$s^{-1/2} \tanh \sqrt{s}$	$G_n(r;0,1) + G_n(r;1/2,1)$
$s^{1/2} \text{csch} \sqrt{s}$	$s^{1/2} \text{sech} \sqrt{s}$	$2(n+1/2)^2\, G_n(r;1/2,3) - G_n(r;1/2,2)$
$s^{1/2} \coth \sqrt{s}$	$s^{1/2} \tanh \sqrt{s}$	$n^2\, G_n(r;0,3) - 1/2 G_n(r,0,2)$ $+(n+1)^2\, G_n(r;1,3) - 1/2 G_n(r;1,2)$
$\text{csch}^2 \sqrt{s}$	$\text{sech}^2 \sqrt{s}$	$4(n+1)^2\, G_n(r;1,2)$

Next $\quad B_n(s) = e^{-2(n+m)\sqrt{s}} / \sqrt{s} = -1/(n+m)\, \{A_n(s)\} \qquad (21)$

or $\quad B_n(r) = 1/(n+m) \cdot 1/r \cdot A_n(r) = G_n(r;m,1) \qquad (22)$

Similarly $C_n(s) = \sqrt{s}\, e^{-2(n+m)\sqrt{s}} = s B_n(s)$

or $\quad C_n(r) = d/d(1/r)\, \{B_n(r)\} = -r^2 (d/dr)\, G_n(r;m,1) \qquad (23)$

which with the aid of equation (18) reduces to

$$C_n(r) = [(n+m)^2 G_n(r;m,3) - 1/2\ G_n(r;m,2)] \qquad (24)$$

Finally $D_n(s) = s^{-(\gamma+1/2)}$

yields $D_n(r) = r^{-\gamma+1/2} / \Gamma(\gamma+1/2) = \rho_\gamma^{-1}\ G(r;0;3-2\gamma) \qquad (25)$

provided γ is a positive integer.

To illustrate the derivation of network parameters in Table 2 from the results of Table 1 consider

$$N(s) = \operatorname{sech} \sqrt{s} = 2/(e^{\sqrt{s}} + e^{-\sqrt{s}}) = 2e^{-\sqrt{s}}\ [1+e^{-2\sqrt{s}}]^{-1} \qquad (26)$$

hence $N_n(s) = 2\ (-1)^n e^{-2(n+1/2)\sqrt{s}}$, where $N(s) = \sum_{n=0}^{\infty} N_n(s) \qquad (27)$

Table 2 yields

$$N_n(r) = (-1)^n (2n+1) G_n(r;1/2,2) \qquad (28)$$

Compare this with $N(s) = \operatorname{csch} \sqrt{s}$

yielding $N_n(r) = (2n+1)\ G_n(r;1/2,2)$

This comparison shows that when $e^{-\sqrt{s}} \rightarrow -e^{-\sqrt{s}}$ in the network function, then $(-1)^n \rightarrow (-1)^{2n} = 1$ in $N_n(r)$.

5. IMPULSE RESPONSE

To interpret the concept of Poisson spectra consider the impulse response for a representative $\overline{RC}$ network. Assuming R = C = 1, the transfer function of a thin film network having an open-circuit termination can be taken as sech $\sqrt{s}$, then

$$\mathcal{L}^{-1}\left\{\operatorname{sech} \sqrt{s}\right\} = \sum_{n=0}^{\infty} (-1)^n (2n+1)\ G_n(r;1/2,2)$$

For each term of the impulse response, alternate maxima and minima occurs for $dG_n/dr = 0$ giving $G_n(\max) = (-1)^n\ 1.85(2n+1)^{-2}$ at $r_n(\max) = 6(2n+1)^{-2}$.

Figure 1 is a plot of terms of the impulse response for n = 0,1,2 with clearly defined properties, identical to those of the Poisson distribution.

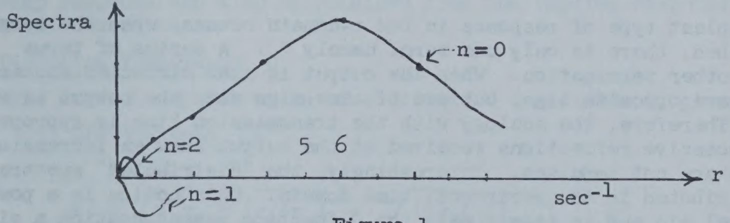


Figure 1

The rapid convergence of the series sketched in Figure 1 clarifies why the Poisson distribution is well suited for r large but why an increasing number of terms are needed for r small. In many applications one or two terms provide an adequate description.

The total impulse response in the r-domain for this network is

$$\sum_{n=0}^{\infty} (-1)^n (2n+1)\ G_n(r;1/2,1)$$

The advantages of expressing the impulse response by G-functions, then, are:

(i) Computational, by providing a formulation in terms of tabulated and well-behaved Poisson distributions.

(ii) Analytic, by providing a systematic procedure which enables one to express the impulse response as Poisson spectra in the reciprocal time domain and which is applicable to all network configurations.

(iii) Conceptual, the physical phenomena governing diffusion can be interpreted by Poisson functions with the same directness as is possible for wave propagation by sinusoidal functions in the direct time domain.

Comparison of $\overline{RC}$ and $\overline{LC}$ Networks

Consider an $\overline{LC}$ transmission line terminated by a resistance R. Assuming L=C=1, the impulse response is obtained from the transfer functions as [5]

$$h(t) = 2R/R+1 . \sum_{n=0}^{\infty} (-1)^n \left(\overline{R-1} / \overline{R+1}\right)^n \delta(t-2n-1)$$

Three terminations of particular importance are:

1. In the matched impedance case R=1, then

$$h(t) = \delta(t-1)$$

2. When R > 1, then

$$h(t) = \sum_{n=0}^{\infty} (-1)^n K_n \ \delta(t-2n-1) \quad , \quad K_n < 1$$

3. When R < 1 but > 0 then the impulse response is

$$h(t) = \sum_{n=0}^{\infty} K_n \ \delta(t-2n-1), \ K_n < 1$$

The actual value of K_n depends upon the termination. Terms of the infinite series for R < 1 and R > 1 describe successive pulses received at the output delayed as the reflections occur at both ends of the transmission line. There is only a single pulse for the matched case and correspondingly a single term in h(t).

In summary, for $\overline{LC}$ networks, the impulse response is equivalent to a spectra of impulses, the magnitude of successive impulses being a function of the termination.

In the $\overline{RC}$ networks, the G-functions represents the reflections for a set of similar terminations. The dissipative character introduces delay in the built-up of energy compared to the instantaneous change of energy for lumped networks under impulse excitation.

The simplest type of response in the r-domain occurs, when the termination is matched, there is only one term, namely G_o. A series of terms results for other termination. When the output is open circuited successive reflections are opposite sign, but are of same sign when the output is short circuited. Therefore, the analogy with the transmission line is appropriate. However, successive reflections received at the output show an increasing spread; they are not impulses. Interestingly, the "distributed" spectra are Poisson distributed in the reciprocal time domain. This notion is a powerful conceptual aid and in itself make the G-function representation a significant analytical tool in the distributed parameter theory.

This analysis has, then, revelaed a very meaningful comparison of $\overline{RC}$ with $\overline{LC}$ networks. This can be summarized as follows:

A. Comparison between wave propagation and diffusion phenomena in distributed structures reveals similarity on account of

- (a) Expansion in terms of successive delays;
- (b) Impulse technique as a tool of analysis;
- (c) The presence of only a single transmission term if a matched termination is provided;
- (d) Alternating series of terms on the basis of transmission and reflection due to the terminating element.

B. The difference is revealed on account of:
- (a) Linear presentation in the direct time domain versus reciprocal time domain.
- (b) Resultant pulses as a series of delta functions versus a series of Poisson functions corresponding to zero spread versus finite spread of successive arrivals of pulses;

6. APPLICATIONS

1. Step and Ramp Response for Thin Film $\overline{RC}$ Network

To demonstrate techniques of obtaining the step response, consider the transfer function

$$N(s) = \operatorname{sech} \sqrt{s} \quad \text{if } R = C = 1$$

Then the output to a step input is

$$V(s) = 1/s \operatorname{sech} \sqrt{s}$$

Using equation (19), one obtains

$$\begin{aligned} V_n(r) &= (-1)^n (2n+1) \int_0^t G_n(r;1/2,2)\,dt \\ &= 2(-1)^n \overline{\left|1/2,(n+1/2)^2 r\right|} \Big/ \overline{\left|1/2\right.} \\ &= 2(-1)^n \operatorname{erfc} \left\{(n+1/2)\sqrt{r}\right\} \end{aligned}$$

The step response is

$$V(r) = \sum_{n=0}^{\infty} V_n(r)$$

Gamma deviates of order 1/2 are normal distributions, hence Incomplete Gamma functions correspond to error functions.

The ramp response can be obtained by numerical evaluation of

$$2(-1)^n \int_{\infty}^{1/r} \overline{\left|1/2,\ (n+1/2)^2\ r\right|}\ d(1/r) \Big/ \overline{\left|1/2\right.}$$

The ramp response can also be obtained from the impulse response function by graphical convolution.

2. Diffusion Transistor

The current amplification of a diffusion transistor is given by [11-14]

$$\alpha(s') = \operatorname{sech} \sqrt{s'}$$

where

$$s' = \sqrt{s\ \tau_w + (\tau_w/\tau_p)} = \sqrt{(s+1/\tau_p)\ \tau_w}$$

Then

$$\alpha_n(r') = (-1)^{n+1}(2n+1)\ G_n(r';1/2,2)$$

Scaling by τ_w and shifting by $1/\tau_p$ gives

$$\alpha_n(r) = 1/\tau_w\ (-1)^{n+1}(2n+1)\ e^{-1/r\tau_p}\ G_n(r\tau_w;1/2,2)$$

The response time of greatest interest occurs during the rise-time, when $r\tau_w$ is between 1 and 10; to estimate the amplitude of the transient, the value of $r\tau_p$ is significant. If $r\tau_p$ exceeds unity, this will heavily dampen the transient during the interval in which $r\tau_w$ is of significance.

3. Tapered $\overline{RC}$ Network

The transfer function of a tapered $\overline{RC}$ network is given by [11-14]

$$N(s) = Re^{c}(s')^{-1/2} \operatorname{csch} \sqrt{s'} \text{ where } s' = c^{2} + sCR$$

The impulse response is then obtained from

$$N_n(r') = 2\sqrt{R}\, e^{c}\, G_n(r';1/2,1)$$

Scaling by RC and shifting by c^2 gives

$$N_n(r) = 2/C \;\; e^{c}\, e^{-\,c^{2}/rRC}\, G_n(rRC;1/2,1)$$

The systematic approach of scaling and shifting in the transform domain in conjunction with a set of computed reference curve for the impulse, brings within reach of the circuit designer a wide range of practical solutions, hitherto inaccessible even conceptually. The use of coupled and interconnected networks in integrated circuits containing chains of interacting networks points towards the need for a procedure to trace transients through complex networks; this is not met by existing techniques. The approach pioneered in these examples offer a radical departure from present procedures and holds promise of being extended to circuits with large numbers of interacting $\overline{RC}$ components.

7. DISCUSSION

Spectra and Convergence

The impulse response of the distributed $\overline{RC}$ networks can be represented as a spectra of modified Poisson distributions in the reciprocal time domain. These spectra are the waves received at the measuring point when an impulse is applied at the sending point. The distributions have exhibited two very important properties. The first is that the spectra alternates in sign if the termination impedance is above the characteristic impedance and it is of same sign if the termination impedance is below the characteristic impedance. The magnitude decreases becuase of dissipation. This behavior is expected if one examines a transmission line, however the fact that the expression for each term of the series is a Poisson distribution in the reciprocal time domain is an interesting result.

Energy Transfer of View

The introduction of the reciprocal time-domain as a vehicle of describing physical phenomena is based in this investigation on pragmatic justification. An epistromological justification of time reversal in physical phenomena requires as axiomatic the probability of energy to concentrate, perhaps similar to the law governing distributed mass in a self-induced gravitational field. A time sequence of signals at the output of an $\overline{RC}$ network will then trigger the first resistive element of the $\overline{RC}$ network to convert heat into voltage across it. Such an event will trigger other capacitors and resistors to accumulate voltage, governed by the law of accumulation requiring charge to flow to effective greatest concentration of charge.

The laws of charge and energy accumulation predict, in fact, that a suitably modified Poisson distribution in the r-domain will tend to concentrate. It remains to show, however, that why will it form a delta function distribution if passed through a $\overline{RC}$ network.

In conclusion, a physical interpretation is always desirable and often possible. The law of charge accumulation of energy concentration or negentropy predicts the trend towards Poisson distributions with decreasing means and decreasing standard deviations, but the trend toward spectral distributions with rapid convergence is a consequence of the type of matching filter used. Negentropy and delay time are not meaningful concepts in lumped $\overline{RC}$ networks, because energy transfer occurs instantly when impulses are applied. It may, therefore, be promising to formulate the differential equations governing charge and energy transfer in the reciprocal time domain. Clearly this cannot simply be done by substituting $r = 1/t$ into the existing differential equations, but requires that these equations be derived on the basis of the laws of charge accumulation, the law of negentropy and probably utilizing the time invariance implied in Maxwell's equations and in conservation laws governing energy and charge.

REFERENCES

1. Weber, E., Linear Transient Analysis, Vol. II, John Wiley, 1956.
2. Bohn, E. V., The Transform Analysis of Linear Systems, Addison Wesley, 1963.
3. Happ, W. W., Synthesis of Distributed RC Networks, IRE Proceedings, pp. 483-484, 1962.
4. Crank, J., The Mathematics of Diffusion, Oxford, Clarendon Press, 1956.
5. Gupta, S. C., Transform and State Variable Methods in Linear Systems, John Wiley and Sons, 1966.
6. Lindholm, F. A. and W. W. Happ, Transient Analysis of Thin Film Structures, Micro-Electronics and Reliability, Vol. 3, pp. 209-226, Pergamon Press 1964.
7. Ghausi, M. S. and Herkowitz, G. D., The Transient Response of Tapered $\overline{RC}$ Networks, IEEE Transactions on Circuit Theory, Vol. CT-10, September 1963.
8. Kelly, J. J. and Ghausi, M. S., On the Dominant Poles of the Distributed RC Network, Journal of Franklin Institute, Vol. 278, No. 2, pp. 108-123, August 1964.
9. Brilliouin, L., Wave Propagation in Periodic Structures, McGraw-Hill, Also Dover Publication, 1953.
10. Seitz, F., The Modern Theory of Sohds, McGraw-Hill, 1940.
11. DePian, L., Linear Active Network Theory, Prentice-Hall, 1940.
12. Carlin, H. J. and Geordano , A.B. , Network Theory , Prentice Hall, 1964.
13. Nisbet, T. R. and Happ, W. W., The Calculus of Deviations Applied to Circuit and Network Analysis, Journal of the British IRE, pp. 437-450, May 1961.
14. Lindholm, F. A. and Hamilton, D. J., Large Signal Models for Junction Transistors, Advances in Solid State Electronics, Vol. VI, 1964.

A FREQUENCY TRANSFORMATION FOR COMMENSURATE TRANSMISSION LINE NETWORKS*

EDWARD G. CRISTAL
Stanford Research Institute
Menlo Park, California

ABSTRACT

The frequency transformation $W = 1/S$, where $S = j \tan \theta$, is investigated for commensurate transmission line networks consisting of stubs, resistors, ideal transformers, and unit elements. This transformation takes transmission line transformers into transmission line lowpass filters, and vice-versa; lowpass (or bandstop) filters into highpass (bandstop) filters, and vice-versa; and elliptic function bandstop filters into elliptic function bandpass filters, and vice-versa. The practicality of the transformation lies in the fact that element values of the transformed network are easily related to the corresponding element values of the original network. The transformation is useful because it provides an alternative viewpoint for synthesis problems, and because it reduces the number of tables of designs needed for various filter types.

INTRODUCTION

Frequency transformations are commonly used in lumped element network theory to convert a given filter network into a related filter network.[1,2] For example, an often used frequency transformation is

$$s' \rightarrow As \quad , \tag{1}$$

where the symbol → stands for "is replaced by," A is a constant; the primed variable is that of the original network, and the unprimed variable is that of the transformed network. The transformation (1) is used to scale the bandwidth of the existing network to another preferred value. Other commonly used frequency transformations in lumped filter theory are[1,2]

$$s' \rightarrow A/s \quad , \qquad \text{(lowpass to highpass transformation)}, \tag{2}$$

$$s' \rightarrow w\left(s/\omega_o + \frac{\omega_o}{s}\right) \quad , \qquad \text{(lowpass to bandpass transformation)}, \tag{3}$$

$$s' \rightarrow \frac{1}{w\left(s/\omega_o + \frac{\omega_o}{s}\right)} \quad , \qquad \text{(lowpass to bandstop transformation)}. \tag{4}$$

It is emphasized that in all cases the usefulness of these transformations lie in the fact that their effects on the responses of the network are easily related to changes in the element values of the network. Because such frequency transformations are available, a given lowpass filter may function as a prototype for a number of different types of filters, making unnecessary the compilation of a multitude of designs for lowpass, highpass, bandpass, and bandstop filters.

Analogous transformations would be equally useful for commensurate transmission-line networks, if they could be developed. For the special class of commensurate transmission-line networks consisting of open- and short-circuited stubs, ideal transformers, and resistors, but without unit elements (i.e., quarter wave-length lines), the transformations (1) through (4) can indeed be used. However, in most cases realizations of commensurate

---

*This work was supported by the U.S. Army Command Laboratories, Fort Mammouth, N.J., under contract DA 28-043-AMC-02266(E).

transmission-line networks without unit elements is impractical or impossible. Unfortunately, in the more general case of commensurate transmission-line networks consisting of open- and short-circuited stubs, ideal transformers, resistors, and unit elements, the usual frequency transformations of lumped-element network theory cannot be used. The fundamental reason for this is that the effects of the transformations on the networks responses are for transmission-line networks with unit-elements, not easily (if at all) related to changes in the element values of the network. An exception to this statement is the frequency transformation*

$$S' \rightarrow 1/S \quad . \tag{5}$$

It has been found that for this transformation the effects on the network elements can be relatively easily accounted for. This transformation corresponds to Eq. (2) with $A = 1$, and possesses corresponding properties. However, its effects on the parameters of transmission line networks can be substantially different from those effects on the parameters of lumped element networks because of the existence of unit elements. The transformation (5) is easily shown to take transmission line filters into transmission line transformers, and vice-versa; lowpass distributed filters into highpass distributed filters, and vice-versa; elliptic function bandstop distributed filters into elliptic function bandpass distributed filters, and vice-versa. The present paper investigates the properties of the transformation (5) and describes the relationships between the original and transformed networks. Several non-trival worked examples using the transformations have been given elsewhere.[4]

THEORY

The general analytical properties of the complex transformation

$$S' \rightarrow 1/S = W \quad , \tag{6}$$

which represents the inversion of the unit circle in the complex plane, are not of particular interest in the present case. The interested reader may find these details in various references.[5,6] Certain specific properties, however, should be pointed out. On the real frequency axis the variable

$$W = 1/S = -j \cot \theta \tag{7}$$

satisfies

$$-j \cot \theta = j \tan (\theta - 90) \quad . \tag{8}$$

Thus, the transformation (6) is equivalent to shifting the origin 90° to the right for all network functions. The result of the translation can be seen to be, for example, that response functions of lowpass (or bandpass) networks become response functions of highpass (or bandpass) networks. This is illustrated by the insertion loss functions shown in Figs. 1(a) and 1(b). Similarly, the response functions of stepped impedance filters are transformed into response functions of stepped impedance transformers; those of short-step transformers[7] are transformed into those of different short-step transformers; and those of bandstop elliptic function filters[8,9,10] are transformed into those of bandpass elliptic function filters.

B.1. Networks of only commensurate unit elements.

We consider first networks consisting of only unit elements. Such networks may represent stepped impedance transformers, lowpass filters, or prototype networks for directional couplers and other filter types.[11,12] Let the original network be referred to as the S'-plane network, and the

---

[*] For commensurate transmission-line networks the variable S represents $j \tan \theta$, where θ is the electrical length of the commensurate transmission lines,[3] and $j = \sqrt{-1}$.

transformed network as the W-plane network. Also let impedances in the S' plane be primed and those in the W plane be unprimed. Herein, for brevity, we present only the results of the theory. Proofs are given in Ref. 4. The relationships between impedances Z_i' in the S' plane and Z_i in the W plane are

$$Z_i = \left[(Z_1')(Y_2')(Z_3') \; . \; . \; . \; (Y_{i-1}') \right]^2 \; Z_i' \quad \text{for i odd,} \tag{9}$$

and

$$Z_i = \left[(Z_1')(Y_2')(Z_3') \; . \; . \; . \; (Z_{i-1}') \right]^2 \; Y_i' \quad \text{for i even.} \tag{10}$$

Equations (9) and (10) are equivalent to formulas for designing half-wave filters from quarter-wave prototype transformers.[13] It has been shown in Ref. 4, however, that these equations are completely general. They apply to any arbitrary cascade of commensurate transmission lines, with the consequence that the network responses in j tan θ are replaced by the same response in -j cot θ.

Let the impedance remaining after the i^th unit element in the W-plane network has been evaluated be $Z_L(W)$. Let the corresponding impedance in the S'-plane network be $Z_L'(S')$. Then

$$Z_L(W) = \frac{\left[(Z_1')(Y_2')(Z_3') \; . \; . \; . \; (Z_i') \right]^2}{Z_L'(W)} \quad \text{for i odd,} \tag{11}$$

and

$$Z_L(W) = \left[(Z_1')(Y_2')(Z_3') \; . \; . \; . \; (Y_i') \right]^2 \; Z_L'(W) \quad \text{for i even.} \tag{12}$$

B.2. Distributed networks of stubs, unit elements, ideal transformers, and resistors.

The transformation theory which was developed for cascaded commensurate unit elements, and summarized in Sect. B.1., has also been extended to networks consisting of stubs, unit elements, ideal transformers, and resistors.[4] The technique for developing the transformed network from the original network follows closely the concepts used for networks of only unit elements. However, again for brevity, only the results of the theory are presented here.

The results of the transformation theory and synthesis procedure shows that in most cases it is unnecessary to perform any detailed mathematical steps. Generally, one can write down the parameters of the transformed network upon inspection of the S'-plane network. The transformation of a given network by the mapping S' → W can be accomplished as follows:

Unit elements in the transformed network (i.e., W-plane network) are developed according to the rules given in Sect. B.1., without regard to the presence of stubs. In the development of stubs in the transformed network there are two cases. 1) If an odd number of unit elements has been developed, shunt open- or short-circuited stubs of the S'-plane network are transformed into series open- or short-circuited stubs, respectively, (and vice-versa) with appropriate immittance scaling. 2) If an even number of unit elements has been developed, shunt open- or short-circuited stubs of the S'-plane network are transformed into series short- or open-circuited stubs, respectively, with appropriate immittance scaling.

"Appropriate immittance scaling" referred to in 1) and 2) above requires the factor (in terms of impedance)

$$\left[Z_1' \; Y_2' \; Z_3' \; . \; . \; . \; Z_i' \right]^2 \quad \text{for i odd,} \tag{13}$$

and

$$\left[Z_1' \, Y_2' \, Z_3' \; . \; . \; . \; Y_i' \right]^2 \qquad \text{for } i \text{ even,} \tag{14}$$

where Z_j' and Y_j' are the impedance and admittance, respectively, of the j^{th} unit element of the S'-plane network.

An example of the results of these rules is afforded by the S'-plane bandstop filter shown in Fig. 2 and the transformed W-plane bandpass filter shown in Fig. 3. The relationships between parameters of the two networks are given below.

$$\begin{aligned}
&H_1 = h_1 \, , \quad H_{12} = h_{12} \, , \quad H_2 = h_{12}^2/h_2 \, , \\
&H_{23} = h_{12}^2/h_{23} \, , \quad H_3 = h_{12}^2 \, h_{23}^{-2} \, h_3 \, , \quad H_{34} = h_{12}^2 \, h_{23}^{-2} \, h_{34} \, , \\
&H_4 = h_{12}^2 \, h_{23}^{-2} \, h_{34}^2/h_4 \, , \quad H_{45} = h_{12}^2 \, h_{23}^{-2} \, h_{34}^2 \, h_{45}^{-1} \, , \\
&H_5 = h_{12}^2 \, h_{23}^{-2} \, h_{34}^2 \, h_{45}^{-2} \, h_5 \, , \quad N^2 = h_{12}^{-2} \, h_{23}^2 \, h_{34}^{-2} \, h_{45}^2 \, , \quad H_6 = h_6
\end{aligned} \tag{15}$$

B.3. Application to distributed elliptic function filters.

The theory presented in Sects. B.1 and B.2 may also be applied to elliptic function filters.[8,9,10] In the general case, elliptic function filters contain unit elements, shunt and series open- and short-circuited stubs, distributed L-C type shunt and series resonant and anti-resonant sections, and distributed non-degenerate Brune sections. Networks of unit elements and stubs have been discussed in Sect. B.2. In the case that the circuit elements of the network are L-C type series and shunt resonant and anti-resonant circuits, the extension of the theory is straight forward. For example, if an even number of unit elements have been developed in the transformed network, a shunt, L-C series resonant circuit will be transformed into another shunt, L-C series resonant circuit with the impedance level appropriately scaled. On the other hand, if an odd number of unit elements have been developed in the transformed network, a shunt, L-C series resonant circuit will be transformed into a series, anti-resonant circuit with the impedance level appropriately scaled. Extension to other types of L-C circuits is obvious.

Application of the theory to Brune sections reveals that a Brune section transforms into another Brune section. Again, there are two cases to consider depending on whether an even or odd number of unit elements have been developed in the transformed network. The transformation of a Brune section after an odd number of unit elements have been developed is shown in Fig. 4(a). In this figure and also Fig. 4(b) the S'-plane Brune section is shown in the T-section form. This is related by elementary transformations[14] to the form using unity coupled coils. The required impedance scaling factor has been neglected in the figure and must be supplied in the actual application. The particular form of the transformed Brune section shown in Fig. 4(a) has been given by Guillemin.[14] Other equivalent circuits are possible, but not presented here.

The transformation of a Brune section after an even number of unit elements have been developed is shown in Fig. 4(b). Again, the impedance scaling factors which have been deleted must be supplied in the actual application.

BANDWIDTH FORMULAS:

The effect of the transformation $S' \rightarrow \frac{1}{S}$ or bandwidth have been determined using Eq. (8). Let θ_2' and θ_1' be the upper and lower band edges of a given network response in the S' plane. Let the fractional bandwidth of the network be defined as

$$w' = 2\,\frac{\theta_2' - \theta_1'}{\theta_2' + \theta_1'}\ . \tag{16}$$

There are two cases to consider:

Case 1: Both θ_2' and θ_1' lie between 0 and 90°.*

Case 2: $\theta_2' = 180^\circ - \theta_1'$.

Case 1: Under the transformation $S' \rightarrow \frac{1}{S}$, the new bandwidth w is given by

$$\frac{1}{w'} + \frac{1}{w} = \frac{90^\circ}{\theta_2' - \theta_1'}\ . \tag{17}$$

Case 2: Under the transformation $S' \rightarrow \frac{1}{S}$ the new bandwidth w is given by

$$w + w' = 2\ . \tag{18}$$

FOUR THEOREMS FOR SYMMETRICAL AND ANTIMETRICAL FILTERS

Define a symmetrical filter as one having impedances which satisfy $Z'_{n+1-i} = r' Z_i'$, where Z_i' is the impedance of the i^{th} stub or unit element normalized to the generator resistance, n is the total number of stub and unit elements in the filter, and r is the ratio of termination to generator resistance.

Define an antimetrical filter as one whose impedances satisfy $Z'_{n+1-i} = \frac{r'}{Z_i'}$. Then, the following four theorems can be stated:

Theorem 1: Under the transformation $S' \rightarrow \frac{1}{S}$ a symmetrical filter having M unit elements, where M is even, goes into another symmetrical filter with respect to the transformed normalized load r. The transformed normalized load r is equal to r if $\frac{M}{2}$ is even, and is equal to $\frac{1}{r}$, if $\frac{M}{2}$ is odd.

Theorem 2: Under the transformation $S' \rightarrow \frac{1}{S}$, a symmetrical filter having M-unit elements, where M is odd, goes into an antimetrical filter with respect to the new normalized load r. The normalized load r satisfies

$$r = \left(Z_1' Y_2' Z_3' \cdots Z'_{\frac{M-1}{2}}\right)^4 \left(Y'_{\frac{M+1}{2}}\right)^2 \tag{19}$$

for $\frac{M+1}{2}$ even, and

$$r = \left(Z_1' Y_2' Z_3' \cdots Y'_{\frac{M-1}{2}}\right) \left(Z'_{\frac{M+1}{2}}\right)^2 \tag{20}$$

for $\frac{M+1}{2}$ odd.

The notation Z_j' represents the j^{th} unit element of the S'-plane network.

Theorem 3: Under the transformation $S' \rightarrow \frac{1}{S}$, an antimetrical filter having M unit elements, where M is odd, goes into a symmetrical filter. The transformed normalized termination r is unity for all cases.

---

*An example of a Case 1 network is the short-step transformer.[6]

Theorem 4: Under the transformation $S' \rightarrow \frac{1}{S}$, an antimetrical filter having M unit elements, where M is even, goes into another antimetrical filter with respect to the new normalized termination r. The normalized termination r satisfies

$$r = \left(Z_1' \, Y_2' \, Z_3' \, . \, . \, . \, Z_{\frac{M}{2}}'\right)^4 \frac{1}{r'}, \tag{21}$$

for $\frac{M}{2}$ odd, and

$$r = \left(Z_1' \, Y_2' \, Z_3' \, . \, . \, . \, Y_{\frac{M}{2}}'\right)^4 r' \tag{22}$$

for $\frac{M}{2}$ even.

It is emphasized that the preceding theorems are based only on the number of unit elements in the S'-plane networks and do not depend in any way on the number of stubs, transformers or resistors.

CONCLUSIONS

A theoretical study has shown that the frequency transformation $S' \rightarrow \frac{1}{S}$, when applied to commensurate transmission line networks of stubs, unit elements, ideal transformers and resistors, may be easily related to changes in impedance values in the network. In most cases, both the form and the element values of the transformed network may be written down by inspection. The transformation will probably be most useful in reducing the number of design tables required for often-used filter designs. However, in some cases it may also be helpful in extending existing tables of designs.

Knowledge of the transformation also provides an alternative viewpoint to various network synthesis problems. Thus, in some network synthesis problems it may prove useful to work in the W plane rather than the S' plane and then transform the resulting network. For example, retention of significant figures in an electronic digital computer may be a problem in synthesizing narrow-band bandpass filters. However, it may not be a problem in synthesizing wide-band bandstop filters. Thus, the latter could be accomplished and the network transformed thereafter.

REFERENCES

1. N.Balbanian, Network Synthesis, pp. 411-414, Prentice Hall Inc., Englewood Cliffs, N.J., 1958.

2. E.A. Guillemin, Synthesis of Passive Networks, pp. 602-607, John Wiley and Sons, New York, 1957.

3. P.I. Richards, "Resistor Transmission Line Circuits," Proc, IRE, Vol. 36, pp. 217-220, February 1948.

4. E. G. Cristal, G. L. Matthaei, Leo Young, "Design of Microwave Filter Networks," Chpt. 2, First Quarterly Progress Report, SRI Project 6025, Contract DA 28-043-AMC-02266(E), Stanford Research Institute, Menlo Park, Calif., (September 1966).

5. R.V. Churchill, Introduction to Complex Variables and Applications, McGraw-Hill Book Company, Inc., N.Y., 1948.

6. W. Kaplan, Advanced Calculus, Chpt. 9, Addism-Wessley Publishing Co., Inc., Reading,Massachusetts, 1957.

7. G.L. Mattahei, "Short-step Chebychev Impedance Transformers," IEEE Trans. on Microwave Theory and Techniques, MTT-14, August 1966.

8. J.O. Scanlan, J.D. Rhodes, "Cascade Synthesis of Distributed Networks," A paper presented at P.I.B. International Symposium on Generalized Networks, April 1966.

9. R. Levy, I. Whitley, "Synthesis of Distributed Elliptic-Function Filters from Lumped-Constant Prototypes," 1966 International Microwave Symposium Digest, pp. 83-89, May 1966.

10. M.C. Horton, R.J. Wenzel, "Realization of Microwave Filters with Equal Ripple Response in both Pass and Stop Band," A paper presented at P.I.B. International Symposium on Generalized Networks, April 1966.

11. G.L. Matthaei, L. Young, E.M.T. Jones, Design of Microwave Filters, Impedance Matching Networks, and Coupling Structures, Chpt. 6, McGraw-Hill Book Co., N.Y., 1964.

12. R. Levy, "Tables of Element Values for the Distributed Low-Pass Prototype Filter," IEEE Trans. on Microwave Theory and Techniques, MTT-13, pp. 514-536, September 1965.

13. G.L. Matthaei, L. Young, E.M.T. Jones, op.cit., pp. 279.

14. E.Z. Guillemin, op.cit., pp. 343-356.

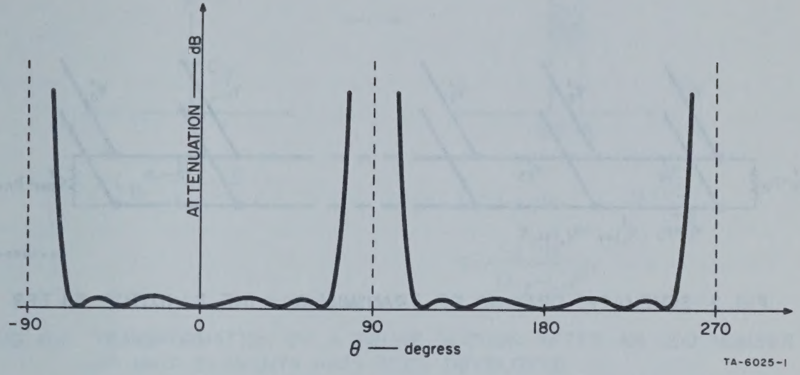


FIG. 1(a) LOWPASS (or Bandstop) ATTENUATION RESPONSE

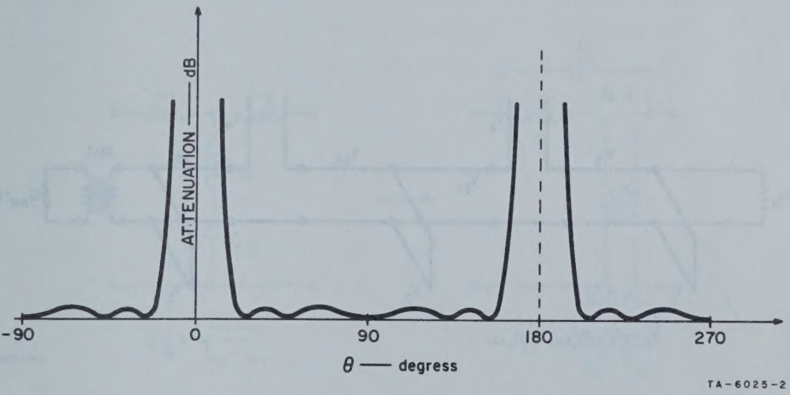


FIG. 1(b) HIGHPASS (or Bandpass) ATTENUATION RESPONSE OBTAINED BY TRANSFORMING THE RESPONSE OF FIG. 1(a) BY THE MAPPING S′ = 1/S

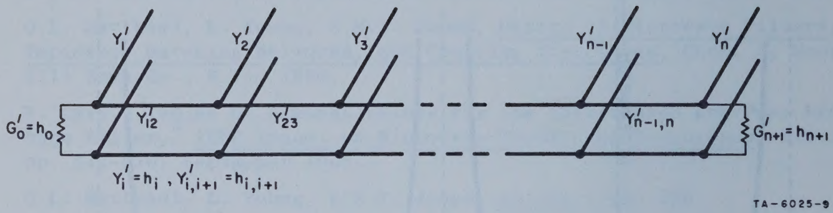

Y_1'
Y_2'
Y_3'
Y_{n-1}'
Y_n'
Y_{12}'
Y_{23}'
$Y_{n-1,n}'$
$G_0' = h_0$
$G_{n+1}' = h_{n+1}$
$Y_i' = h_i$, $Y_{i,i+1}' = h_{i,i+1}$
TA-6025-9

FIG. 2 SCHEMATIC DRAWING OF TRANSMISSION-LINE BANDSTOP FILTER

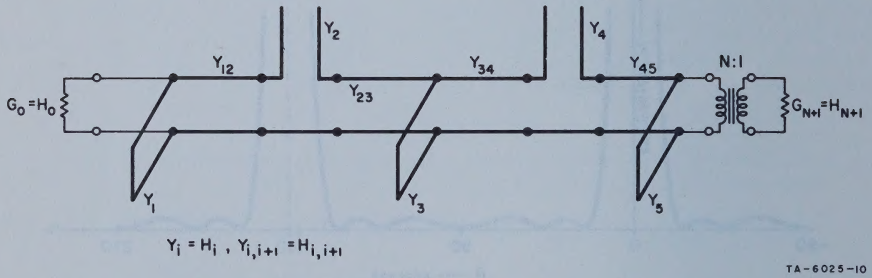

Y_2
Y_4
Y_{12}
Y_{34}
Y_{45}
N:1
Y_{23}
$G_0 = H_0$
$G_{N+1} = H_{N+1}$
Y_1
Y_3
Y_5
$Y_i = H_i$, $Y_{i,i+1} = H_{i,i+1}$
TA-6025-10

FIG. 3 TRANSMISSION-LINE BANDPASS FILTER DEVELOPED FROM A BANDSTOP FILTER USING THE FREQUENCY TRANSFORMATION S′ = 1/S

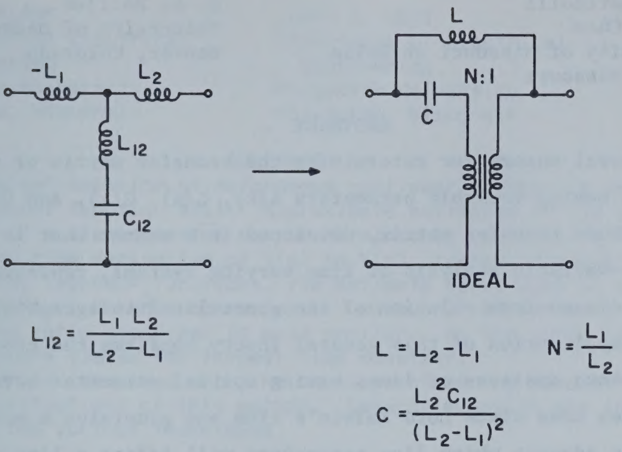

-L₁
L₂
L₁₂
C₁₂
L
N:1
C
IDEAL
$L_{12} = \frac{L_1 L_2}{L_2 - L_1}$
$L = L_2 - L_1$
$N = \frac{L_1}{L_2}$
$C = \frac{L_2^2 C_{12}}{(L_2 - L_1)^2}$

FIG. 4(a) TRANSFORMATION OF A BRUNE SECTION AFTER AN ODD NUMBER OF UNIT ELEMENTS HASS BEEN DEVELOPED
(impedance scaling has been neglected)

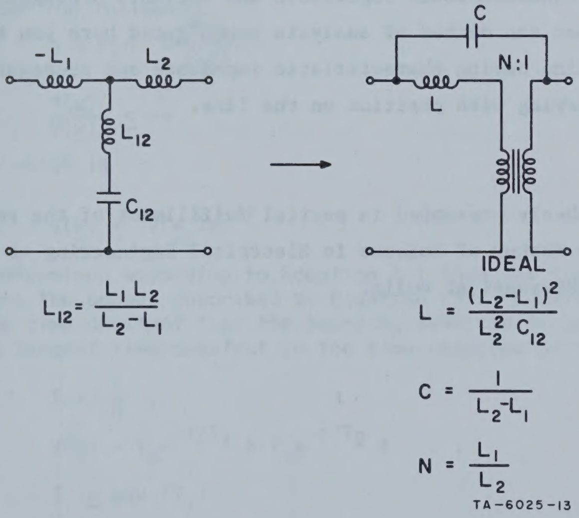

-L₁
L₂
L₁₂
C₁₂
C
N:1
L
IDEAL
$L_{12} = \frac{L_1 L_2}{L_2 - L_1}$
$L = \frac{(L_2 - L_1)^2}{L_2^2 C_{12}}$
$C = \frac{1}{L_2 - L_1}$
$N = \frac{L_1}{L_2}$
TA-6025-13

FIG. 4(b) TRANSFORMATION OF A BRUNE SECTION AFTER AN EVEN NUMBER OF UNIT ELEMENTS HAS BEEN DEVELOPED
(impedance scaling has been neglected)

Variable Parameter Transmission Line Theory

E. C. Bertnolli
Yu Kuo Chen
University of Missouri at Rolla
Rolla, Missouri

C. A. Halijak
University of Denver
Denver, Colorado

ABSTRACT

The general theory for determining the transfer matrix of a transmission line having variable parameters R(x), L(x), G(x), and C(x) is developed. This transfer matrix, developed in a manner that is analogous to the state-variable analysis of time varying systems, represents the first exact closed-form solution of the generalized telegraphist's equations. Application of this general theory confirms the transient and steady state analyses of lines having special parameter variations that have been made since Lord Kelvin's time and generates a method for specifying in advance which line parameters will affect a line's performance.

The literature on variable parameter lines is limited to those lines which have propagation constants that are identical at each point on the line and characteristic impedances which vary with position along the line. This paper includes an example of a new class of lines which have constant characteristic impedances and variable propagation constants. It is shown that the method of analysis promulgated here can be used readily on a line having characteristic impedance and propagation constant, each varying with position on the line.

Portion of a thesis presented in partial fulfillment of the requirements for the degree Master of Science in Electrical Engineering at the University of Missouri at Rolla.

ESTIMATING THE TRANSIENT RESPONSE OF NONLINEAR SYSTEMS

GARY L. BLANK
Department of Electrical Engineering
Marquette University
Milwaukee, Wisconsin

LARRY J. LEVY
Department of Electrical Engineering
Marquette University
Milwaukee, Wisconsin

ABSTRACT

The transient behavior of autonomous nonlinear systems is examined by using the Lyapunov function $V(\underline{x})$. Approximate estimates of the dominant time constant of a stable system can be obtained by considering the ratio of the negative time derivative of $V(\underline{x})$ to $V(\underline{x})$ itself. Since a system may possess many Lyapunov functions, the estimate based upon an arbitrary $V(\underline{x})$ will in general be only approximate. However, the estimate can be improved if the ratio described is made smaller. As the ratio is minimized the estimate yields the largest time constant.

This paper is concerned with the development of the theoretical aspects and applications of this method. Several examples are included to illustrate the various techniques.

INTRODUCTION

The Lyapunov function $V(\underline{x})$ is a nonlinear function of the state variables of a system. It can be regarded as a "position" vector in the state space and its time derivative $\dot{V}(\underline{x})$ is a measure of the time rate of change of this vector. If the system is asymptotically stable in some region and $V(\underline{x})$ is positive-definite, then $\dot{V}(\underline{x})$ is negative-definite implying that $V(\underline{x})$ is approaching a point of stable equilibrium. The smaller the rate of decrease the less stable is the system in a qualitative sense. In order to establish a lower bound estimate on the rate of decrease, consider the following.

Let
$$\eta \equiv \min_{\underline{x}} \left\{\frac{-\dot{V}(\underline{x})}{V(\underline{x})}\right\} \tag{1}$$

then
$$\frac{\dot{V}(\underline{x})}{V(\underline{x})} \leq -\eta \tag{2}$$

the solution of which is
$$V(\underline{x}) \leq V(\underline{x}_o)e^{-\eta t} \tag{3}$$

If η is determined according to Equation (1) then the time response of $V(x)$ is within the bounds described by Equation (3). Equation (3) also implies that the time constant T of the bounding function is greater than or equal to the largest time constant in the time response of $V(\underline{x})$ where
$$T \equiv \frac{1}{\eta}, \tag{4}$$
$$V(\underline{x}) = V_o e^{-t/T_1} + V_1 e^{-t/T_2} + \ldots \tag{5}$$

and
$$T \geq \max(T_i)$$

Since $V(x)$ is generally a quadratic function of the state variables then T is half the magnitude of the time constant τ_i of the individual state variables.

Let
$$\tau \equiv \max(\tau_i) \tag{7}$$

then
$$T \geq \frac{\tau}{2} \tag{8}$$

or $$\tau \leq \frac{2}{\eta} \tag{9}$$

Inequality (9) and Equation (1) provide the fundamental relation between the system dominant time constant τ and the Lyapunov function. Since $V(\underline{x})$ is not unique, η will only be an estimate which becomes more accurate as $-\dot{V}(\underline{x})/V(\underline{x})$ is minimized.

LINEAR SYSTEMS

For the autonomous system described by

$$\dot{\underline{x}} = A\underline{x} \tag{10}$$

the quadratic form of $V(\underline{x})$ is

$$V(\underline{x}) = \underline{x}^{\dagger}P\underline{x} \tag{11}$$

where $\underline{x}^{\dagger}$ is the transpose of $\underline{x}$ and P is a positive-definite, real, and symmetric matrix.

Differentiating Equation (11) with respect to time,

$$\dot{V}(\underline{x}) = \underline{x}^{\dagger}P\,\dot{\underline{x}} + \dot{\underline{x}}^{\dagger}P\,\underline{x} \tag{12}$$

Substituting Equation (10) into Equation (12)

$$\dot{V}(\underline{x}) = \underline{x}^{\dagger}[PA + A^{\dagger}P]\,\underline{x} \tag{13}$$

Let $$-B = PA + A^{\dagger}\,P \tag{14}$$

then $$\dot{V}(\underline{x}) = -\underline{x}^{\dagger}B\,\underline{x} \tag{15}$$

If B is positive-definite the system is asymptotically stable in the large.

Combining Equations (1), (11), and (15)

$$\eta = \min_{\underline{x}} \left\{\frac{-\dot{V}(\underline{x})}{V(\underline{x})}\right\} = \min_{\underline{x}} \left\{\frac{\underline{x}^{\dagger}B\,\underline{x}}{\underline{x}^{\dagger}P\,\underline{x}}\right\} \tag{16}$$

Since P is real, symmetric, and positive-definite, it follows that [Reference 1]

$$\eta = \min \lambda_i\,(BP^{-1}) \tag{17}$$

where $\lambda_i\,(BP^{-1})$ is a characteristic value of the matrix (BP^{-1}) and is determined from

$$|BP^{-1} - \lambda I| = 0 \tag{18}$$

Without Equations (17) and (18) η could have been obtained by considering partial derivatives of the ratio of the polynomials in Equation (16) with respect to the state variables.

Example 1

Consider

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}-4 & 4\\ 2 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} \tag{19}$$

Choosing

$$P = \begin{bmatrix}7/40 & 4/40\\ 4/40 & 6/40\end{bmatrix} \tag{20}$$

then

$$-B = PA + A^{\dagger}P = -I \tag{21}$$

and

$$\frac{-\dot{V}(\underline{x})}{V(\underline{x})} = \frac{\underline{x}^{\dagger}\underline{I}\,\underline{x}}{\underline{x}^{\dagger}P\,\underline{x}} = \frac{40(x_1^{\,2} + x_2^{\,2})}{7x_1^{\,2}+8x_1x_2+6x_2^{\,2}} \tag{22}$$

If the partial derivatives of Equation (22) with respect to x_1 and x_2 are each set equal to zero to minimize the ratio of polynomials, considerable computation effort yields

$$\eta = \min_{\underline{x}} \left\{\frac{-\dot{V}(\underline{x})}{V(\underline{x})}\right\} = 3.8 \tag{23}$$

From Equation (9),

$$\tau \leq 0.527 \tag{24}$$

If τ had been calculated by the minimum eigenvalue method described by Equations (17) and (18), then

$$\eta = \min \lambda_i \ (B - \lambda P) \tag{25}$$

Solving the characteristic equation **yields** the eigenvalues

$$\lambda_1 = 16.2, \text{ and } \lambda_2 = 3.8$$

Thus,

$$\tau \leq \frac{2}{3.8} = 0.527 \tag{26}$$

which is the same result as before but computationally easier to obtain in this case.

The largest actual time constant can be determined from the A matrix for this simple example. The eigenvalues of A are $\lambda_1 = 2$, and $\lambda_2 = 8$. The time constant in question is 0.5 indicating that the previous result is about 6% larger than the exact result.

NONLINEAR SYSTEMS

For the nonlinear autonomous system described by

$$\dot{\underline{x}} = f(\underline{x}) \tag{27}$$

a development according to Krasovskii [Reference 2] uses

$$V(\underline{x}) = \underline{f}^{\dagger}(\underline{x})\ A\ \underline{f}(\underline{x}) \tag{28}$$

where $A = A^{\dagger}$ is real, constant, and positive-definite. Then

$$\dot{V}(\underline{x}) = \underline{f}^{\dagger}(\underline{x})\ A\ \dot{\underline{f}}(\underline{x}) + \dot{\underline{f}}^{\dagger}(\underline{x})\ A\ \underline{f}(\underline{x}) \tag{29}$$

It can easily be shown that

$$\dot{\underline{f}}(\underline{x}) = J\ \dot{\underline{x}}, \text{ and } \dot{\underline{f}}^{\dagger}(\underline{x}) = \dot{\underline{x}}^{\dagger}J^{\dagger} \tag{30}$$

where J is the Jacobian matrix given by

$$J = \begin{bmatrix} \frac{\partial f_1(\underline{x})}{\partial x_1} & \frac{\partial f_1(\underline{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\underline{x})}{\partial x_n} \\ \frac{\partial f_2(\underline{x})}{\partial x_1} & & & \\ \vdots & & & \\ \frac{\partial f_n(\underline{x})}{\partial x_1} & \cdots & & \frac{\partial f_n(\underline{x})}{\partial x_n} \end{bmatrix} \tag{31}$$

Equation (29) now becomes

$$\begin{aligned} \dot{V}(\underline{x}) &= \underline{f}^{\dagger}(\underline{x})\, A\, J\, \underline{\dot{x}} + \underline{\dot{x}}^{\dagger}\, J^{\dagger}\, A\, \underline{f}(\underline{x}) \\ &= \underline{f}^{\dagger}(\underline{x})\, [A\, J + J^{\dagger} A]\, \underline{f}(\underline{x}) \end{aligned} \tag{32}$$

Let

$$-\hat{J}(\underline{x}) = A\, J + J^{\dagger} A \tag{33}$$

then

$$\dot{V}(\underline{x}) = -\underline{f}^{\dagger}(\underline{x})\, \hat{J}(\underline{x})\, \underline{f}(\underline{x}) \tag{34}$$

If $\hat{J}(\underline{x})$ is positive-definite for a certain region of $\underline{x}$ then the system is asymptotically stable in this region. From Equations (1), (28), and (34),

$$\eta = \min_{\underline{x}} \left\{ \frac{\underline{f}^{\dagger}(\underline{x})\, \hat{J}(\underline{x})\, \underline{f}(\underline{x})}{\underline{f}^{\dagger}(\underline{x})\, A\, \underline{f}(\underline{x})} \right\} \tag{35}$$

The minimum value of λ which satisfies the characteristic equation

$$\left| \hat{J}(\underline{x}) - \lambda A \right| = 0 \tag{36}$$

provides the value of η of Equation (35) [Reference 4].

Example 2

Consider the system described by

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 - x_2 - a x_1^{3} \end{aligned} \tag{37}$$

A Lyapunov function is given by

$$V(\underline{x}) = \underline{f}^{\dagger}(\underline{x})\, A\, \underline{f}(\underline{x}) \tag{38}$$

where

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{39}$$

Since

$$J = \begin{bmatrix} 0 & 1 \\ -1-3ax_1^{2} & -1 \end{bmatrix} \tag{40}$$

then

$$\hat{J}(\underline{x}) = \begin{bmatrix} 2 + 6ax_1^{2} & 1 + 6ax_1^{2} \\ 1 + 6ax_1^{2} & 2 \end{bmatrix} \tag{41}$$

This system is asymptotically stable only when

$$2 + 6ax_1^2 > 0 \tag{42}$$

and

$$4 + 12\,ax_1^2 - (1 + 6ax_1^2)^2 > 0 \tag{43}$$

Inequalities (42) and (43) are satisfied for $0 \leq x_1 < \frac{\pm 1}{4\sqrt{3}\sqrt{2a}}$

The characteristic equation (36) can now be solved to yield

$$\lambda(\underline{x}) = 1 \pm 2\sqrt{3}\;a\,x_1^2 \tag{44}$$

For $a = 1$, $\lambda(\underline{x})$ is minimum for $x_1 = 0$. Thus

$$\eta = \lambda_{min} = 1 \tag{45}$$

The result of Equation (45) will now be compared with that obtained by applying the method of Vogt [Reference 3]. In both cases stability is assured in the neighborhood of the origin. This is observed in Inequalities (42) and (43), and by the linearization technique of Vogt.

The system equation (37) may be written as

$$\dot{\underline{x}} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}\underline{x} + \begin{bmatrix} 0 \\ -ax_1^3 \end{bmatrix}$$

$$= G\underline{x} + \underline{g}(\underline{x}) \tag{46}$$

Consequently,

$$\lambda_i\,(G) = -\frac{1}{2} \pm j\frac{\sqrt{3}}{2} \tag{47}$$

where the eigenvalues are considered for the linear portion (matrix G) of Equation (46). Choosing

$$P = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{48}$$

then

$$-B = G^{\dagger}\,P + P\,G$$

$$= \begin{bmatrix} -2 & -1 \\ -1 & -2 \end{bmatrix} \tag{49}$$

Equivalently,

$$B = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \tag{50}$$

Vogt has shown that solving the characteristic equation

$$|B - \lambda\,P| = 0 \tag{51}$$

for the minimum λ_i is identical to evaluating the expression

$$\min_{\underline{x}} \left\{\frac{\underline{x}^{\dagger}\,B\,\underline{x}}{\underline{x}^{\dagger}\,P\,\underline{x}}\right\} \tag{52}$$

From Equation (51) $\lambda_{min} = 1 = \eta$.

This result compares favorably with that of Equation (45) which was obtained by Krasovskii's method. However, Krasovskii's method does not have to be restricted to a neighborhood of the origin, but can be used to estimate the transient response for larger perturbations.

EXTENSION OF KRASOVSKII'S METHOD

An extension of Krasovskii's method is based upon

$$V(\underline{x}) = \underline{x}^{+} P(\underline{x})\, \underline{x} \tag{53}$$

and

$$\dot{V}(\underline{x}) = -\underline{x}^{+} B(\underline{x})\, \underline{x} \tag{54}$$

Here, however, $V(\underline{x})$ is not restricted as in Krasovskii's method. In using Krasovskii's method to find the time constant, a particular Lyapunov function had to be used, namely, a $V(\underline{x})$ function which satisfied Equations (28) and (34). In this extension of Krasovskii's method $V(\underline{x})$ may be any suitable Lyapunov function which must now be reformulated to conform with Equations (53) and (54). There are several $\lambda_i(\underline{x})$ which satisfy the characteristic equation

$$|B(\underline{x}) - \lambda(\underline{x})\, P(\underline{x})| = 0 \tag{55}$$

Of these, the particular $\lambda_i(\underline{x})$ which passes closest to the origin in the λ space is evaluated at the minimal point.

Example 3

Consider the system described by

$$\dot{x}_1 = x_2 \tag{56}$$

$$\dot{x}_2 = -x_1 - x_2 - ax_1^3$$

This system is identical to that of Example 2. Both $V(\underline{x})$ and $\dot{V}(\underline{x})$ can be reformulated to conform with Equations (53) and (54). Thus

$$P(\underline{x}) = \begin{bmatrix} 2(1 + ax_1^2)^2 & 1 + ax_1^2 \\ 1 + ax_1^2 & 2 \end{bmatrix} \tag{57}$$

and

$$B(\underline{x}) = \begin{bmatrix} 2(1 + ax_1^2)^2 & -(1+ax_1^2)(ax_1^2-1) \\ -(1 + ax_1^2)(ax_1^2-1) & 2 - 6ax_1^2 \end{bmatrix} \tag{58}$$

Using Equation (55)

$$\lambda(\underline{x}) = 1 \pm 2\sqrt{3}\; ax_1^2 \tag{59}$$

which yields the same result as Example 2.

CONCLUSION

The examples given illustrate how the transient response of stable systems can be estimated. In particular, Examples 2 and 3 were concerned with nonlinear systems. The three techniques described are based upon the Lyapunov function. The first is the simple case of a linear system. The second is based upon the development of Krasovskii. The third method is a modification of the second the main difference being that in Krasovskii's method there were restrictions on the $V(\underline{x})$ function. These restrictions are removed if any permissible $V(\underline{x})$ function is reformulated to yield a $P(\underline{x})$ matrix and a $B(\underline{x})$ matrix.

ACKNOWLEDGEMENT

The authors wish to thank H. Messinger for her invaluable help in the preparation of the manuscript.

REFERENCES

1. Kalman, R. E.,and Bertram, J. E., "Control System Analysis and Design via the Second Method of Lyapunov." *Transactions of the ASME*, Series D, 82 (1960), 371.

2. Gibson, J. E., *Nonlinear Automatic Control*, McGraw-Hill Book Co., New York, New York, 1963.

3. Vogt, W. G., "Transient Response from the Lyapunov Stability Equation," *Proceedings of the JACC*, (1965), 23.

4. Cunningham, W. J., "An Introduction to Lyapunov's Second Method", *Work Session in Lyapunov's Second Method*, edited by L. F. Kazda, (1960), 1.

PERFORMANCE INVESTIGATION OF NONLINEAR SYSTEMS BY OBSERVING EIGENVALUE MOVEMENT

R. WASTA
Department of Electrical Engineering
Pennsylvania State University
University Park, Pennsylvania 16802

S.A. STEELE
General Electric Company
Space Technology Center
Valley Forge, Pa.

ABSTRACT

The question has often been asked, "How far can eigenvalues move in a nonlinear system and still give useful information in the analysis". This paper attempts to give some insight into this problem if the eigenvalue movement is properly interpreted. A method of analysis is proposed and examined for a wide class of nonlinear systems subject to step drives. The class predominately contains third order systems, but a fourth order and seventh order system are investigated. The purpose of the performance investigation is to indicate whether the nonlinear system response is oscillatory or exhibits no overshoot. While more than one nonlinearity may be present and affect the response greatly, the nonlinearities are of a "smooth" nature. Classically, the eigenvalue locations are useful only in linear systems. However, efforts have been made to extend the concept to nonlinear cases. Determination of oscillatory or aperiodic motion by observing root (eigenvalue) excursions has been done for second order systems. It is the extension of the root excursion method to higher order problems that is the major purpose of this paper. A characteristic equation associated with the nonlinear differential equation of the system is investigated. The roots of this defined characteristic equation vary as a function of the dependent variable. The proper way to evaluate the root movement to determine performance (not stability) is discussed. The results are applicable primarily to systems with intentional nonlinearities, added to improve response. Inherent in this technique is the fact that the number of nonlinearities is not critical. Procedures indicating whether a third order nonlinear system can be analyzed by this method are given. On third order systems the method may be applied in a necessary and sufficient way. The procedure as outlined gives sufficient type results for the higher order systems investigated.

SYSTEM PERFORMANCE OF CONTROL SYSTEMS WITH TWO NONLINEARITIES

C. A. PELLEGRINI H. M. YOCKEY G. J. THALER
U. S. Naval Postgraduate School,
Monterey, California

ABSTRACT: Parameter plane techniques and describing function theory are combined to obtain a dynamic describing function curve which represents the movement of the operating or $M(\alpha, \beta)$ point on the parameter plane. When the system characteristic equations fall into two general forms, the transient response as well as relative stability can be established under certain restrictions. Specific examples are analyzed and the theoretical results compared with analog computer simulation.

INTRODUCTION

The introduction of coefficient plane techniques by Mitrovic[1] provided the analyst or synthesizer of servo systems with control of the first two coefficients of the linear system characteristic equation. This technique was later generalized by Siljak[2] in which any pair of coefficients of the characteristic equation could be varied. Subsequent work by Siljak[3] using parameter plane techniques provided the capability of controlling a servo system by varying any two variables which could be components of any coefficient of the characteristic equation.

Parameter plane techniques have been used to predict the relative stability of nonlinear systems[4]. Additionally, it has been shown[5,6] for special cases that the transient performance of systems containing two nonlinear elements can be predicted when describing function theory is combined with parameter plane techniques. This paper will show that for three broad classes of nonlinear systems the relative stability can be predicted accurately and also that the transient response can be predicted under certain conditions.

CLASSIFICATION OF NONLINEAR SYSTEMS

As developed by Siljak[3], the general coefficient, a_k, of the characteristic equation, i.e., $f(s) = \sum_{k=0}^{n} a_k s^k$, can take on two general forms shown below which are applicable to systems containing two nonlinear elements.

$$\text{(a)} \quad a_k = b_k\alpha + c_k\beta + d_k \qquad \text{(b)} \quad a_k = b_k\alpha + jc_k\beta + d_k \tag{1}$$

If the a_k terms are of either form where the α and β variables are nonlinear elements or components of nonlinearities, parameter plane techniques may be used. Based on this formulation nonlinear systems which are subject to an initial excitation can be conveniently classified into three general classes. Figure 1 depicts the three cases which satisfy the requirements of the general coefficient in (1).

In each case the letter X denotes an initial condition excitation which for expository purposes will be considered as the peak value of a sinusoidal signal. Since describing function theory will be used, it will be assumed that sufficient low frequency filtering is provided by the system.

Case 1 denotes those systems for which both nonlinear elements have the same signal excitation. For Case 2 the two nonlinear elements are related to the initial signal excitation by a linear differential equation. In Case 3 the nonlinear elements are interdependent in that the input to either nonlinear element is affected by the characteristics of the other.

DYNAMIC DESCRIBING FUNCTION CONCEPT

When the nonlinear system fits one of the three forms delineated, a net of $\zeta - \omega_n$ curves can be computer-plotted because of the recursive nature of the Chebyshev polynomial expansion of equations (1). When using describing function theory, constant ω curves are desired where $\omega = \omega_n \sqrt{1-\zeta^2}$. If the complex frequency $s = -\zeta\omega_n \pm j\omega_n \sqrt{1-\zeta^2}$ is expressed as $s = -\sigma \pm j\omega$, constant σ or constant ω curves may be obtained because the characteristic equation $f(s) = \sum_{k=0}^{n} a_k s^k$ can be expanded in terms of binomial coefficients (Siljak).

Since the α and β coordinates of the parameter plane represent the nonlinearities in the system, the describing functions of these elements may be used to obtain a net of amplitude, A, versus constant ω curves. Scaling the coordinates for both the constant ζ-ω or σ-ω curves on the parameter plane identically to the A-ω grid allows one to place one set of curves atop the other to find the common radian frequencies, ω, which satisfy both plots. The concomittant amplitude, A, of the initial signal excitation is also available. The series of points obtained in this fashion is called dynamic describing function locus because it represents the movement of the operating or $M(\alpha, \beta)$ point on the parameter plane. By connecting these points by a smooth curve the amplitude and system frequencies under limit cycle or oscillatory conditions may be predicted at the point of crossing the $\zeta = 0$ or $\sigma = 0$ curve. That portion of the locus enclosed by $0 \le \zeta \le 1$ is the stable region.

TRANSIENT RESPONSE

The transient response of a nonlinear system may be predicted from the system's dynamic describing function curve if the conditions shown in the mathematical development[7] are met. In addition to the dominancy requirements the variations of ζ and ω must change slowly with time. That is, if the dynamic describing function curve lies in areas of rapidly changing ζ and ω, the accuracy with which the transient response is predicted will deteriorate.

In brief, the method is applicable to systems in which describing functions are valid representations of the nonlinear elements. The transient response predictions are accurate when the dynamic describing function curve indicates that a pair of complex roots is dominant, and the accuracy of these predictions deteriorates as the contribution of non-dominant roots become more important. Some specific examples will demonstrate the technique and the results obtainable:

Case 1 Two nonlinear elements with a common signal excitation.

Consider the block diagram in Figure 1a where the transfer functions are identified as follows:

$$G_1(s) = 1 \qquad G_3(s) = 1$$

$$G_2(s) = \frac{10}{s^2} \qquad G_4(s) = \frac{s}{s+10} \tag{2}$$

The characteristic equation for this system is

$$f(s) = s^3 + 10s^2 + (10N_1 + 10N_2)s + 100N_1 = 0 \qquad (3)$$

where N_1 and N_2 are considered as gain variable nonlinearities. By letting $\alpha = N_1$ and $\beta = N_2$ a net of ζ-ω or σ-ω parameter plane curves may be plotted. For comparison purposes the effect of changing N_1 and N_2 may be observed by noting the changing dynamic describing function loci obtained on the constant ζ-ω curves. For this example N_1 and N_2 are considered as dead zone elements. The dynamic describing function curve is very easy to obtain: choose a sequence of input signal amplitudes, evaluate $N_1 = \alpha$ and $N_2 = \beta$ for each amplitude and plot on the parameter plane of Figure 2. Three cases are considered, namely, (1) dead zone width of $N_1 > N_2$, (2) dead zone width of $N_2 = N_1$, and (3) the dead zone width of $N_1 < N_2$. Figure 2 shows the dynamic describing function locus for the three conditions enumerated above. Some typical expected signal excitation amplitudes are shown adjacent to the curves. Curves 2 and 3 are in relatively slowly changing regions of ζ and ω but part of curve 1 lies in a region where ζ and ω change rapidly with time. The transient response for Curve 3 is shown in Figure 3. The calculated transient response agrees closely with the analog computer results shown in the upper right hand corner. For the reasons already mentioned the transient performance responses for Curves 1 and 2 are not as good.

Case 2 Two nonlinear elements related to a common signal excitation by a linear differential equation.

For this example the block diagram in Figure 1b has the following transfer functions:

$$G_1(s) = 1 \qquad G_3(s) = 1$$
$$G_2(s) = \frac{20}{s(s+1)(s+2)(s+4)} \qquad G_4(s) = s \qquad (4)$$

The system characteristic equation is

$$f(s) = s^4 + 7s^3 + 14s^2 + (8 + 20N_2)s + 20N_1 = 0 \qquad (5)$$

The nonlinear elements, N_1 and N_2 are considered as dead zone and saturation elements respectively. In this case the input signal to N_2 is frequency sensitive due to the tachometer feedback. Plotting the variations of the describing functions for N_1 and N_2 obtains a typical A-ω grid shown in Figure 4. Figure 5 shows two dynamic describing function loci plotted on the σ-ω parameter plane curves. For both curves the saturation level of N_2 was fixed at ±1 volt and the dead zone width of N_1 at ±1 volt and ±3 volts in curves 1 and 2 respectively. The dotted lines indicate several of the constant ω curves corresponding to the A-ω grid in Figure 4. Figures 6 and 7 show two typical calculated transient responses with the analog computer traces in the upper right hand corner. The correspondence is reasonably good except that a real root effect is apparent in the computer traces. As previously mentioned the graphical technique used to predict the approximate transient response is limited to second order dominancy conditions.

Case 3 Two nonlinear elements related to a common signal excitation by a nonlinear differential equation.

In this case the input signal to the nonlinear elements depends to some degree on the characteristics of the preceding nonlinear element. For an example of this case consider the transfer functions for Figure 1c as follows:

$$G_1(s) = \frac{10}{s + 1} \qquad G_2(s) = \frac{4K}{s(s + 2)} \tag{6}$$

K is a variable gain element and the nonlinear elements N_1 and N_2 are saturation and backlash respectively. The characteristic equation for this system becomes

$$f(s) = s^3 + 3s^2 + 2s + 40KN_1(N_a + jN_b) = 0 \tag{7}$$

where the nonlinearity N_2 has been expressed in terms of its real component, N_a and imaginary component, N_b. Since the coefficients of equation (7) are of the form in equation 1(b), we can let $\alpha = N_1N_a$ and $\beta = N_1N_b$. Because of the nature of the backlash element, the system will have a limit cycle. The amplitude and frequency of the limit cycle can be predicted as before by plotting a grid of constant A-ω curves and noting the intersection with the ζ=0 curve of the parameter plane. For analysis and simulation purposes the saturation level was set at ±1 volt and the backlash width to ±1 volt. Figures 8 and 9 show the theoretical plots to obtain the amplitude and frequency of the limit cycle for K = .15 and K = .25 respectively. The results are compared below.

	Computer		Theoretical	
	Amplitude	ω	Amplitude	ω
K = .15	.23 volts	.52	.3 volts	.49
K = .25	.45 volts	.68	.55 volts	.65

Part of the discrepancy between the radian frequencies and signal amplitudes is that in describing function theory the radian frequency is a second order approximation while amplitude is a first order approximation. Keeping these limitations in mind, the results in the above table correlate quite well.

CONCLUSIONS

When properly applied the methods described in this paper predict the absolute stability of a system with two nonlinear elements, and also predict (with reasonable accuracy) the amplitude and frequency of any limit cycles. When the system is stable, accurate predictions of transient response are obtained if second order dominance conditions exist; less accurate transient predictions are obtained when a pair of complex roots is not clearly dominant.

APPENDIX

The mathematical justification of the graphical response technique used in this paper appears in Siljak's paper[7]. Briefly the method depends on the Krylov-Bogoliubov generalization for calculating the transient response for second order systems. When second order dominance is justified, the method can be extended to higher order systems if certain restrictions in the Krylov-Bogoliubov solution are met such as ζ and ω_n varying slowly with time.

The technique assumes that the solution of a second order nonlinear system may be expressed as

$$X = A(t) \sin \phi \tag{1}$$

where the signal amplitude, A, and phase, ϕ, are given as

$$\frac{dA}{dt} = -\zeta\omega_n A = -\sigma A \qquad \frac{d\phi}{dt} = \omega_n \sqrt{1-\zeta^2} = \omega \tag{2}$$

This pair of equations represents the exponential decay of the signal envelope and the describing function frequency respectively. The variables, ζ and ω_n, which satisfy the above pair of equations also satisfy the parameter plane equations. However, it is simpler to use the σ and ω variables because needless conversions between the ω_n and ω are eliminated.

The dynamic describing function places in evidence the characteristic roots of the nonlinear system at each instant of time. To graphically plot the transient response a particular initial signal, A_o, is selected on the dynamic describing function curve and the value of σ and ω read at that point. The signal is assumed to decay exponentially for a short distance along the straight line approximation, terminated by $1/\sigma$, the decay time. At this point a new value of σ and ω, corresponding to the amplitude at the termination of the previous step are read from the parameter plane. The procedure determines the signal envelope after several iterations. The zero signal crossings are determined by plotting the value of ω read from the parameter plane against the values of time corresponding to the amplitude versus time envelope. By graphical integration, i.e., $\phi = \int \omega dt$, the zero crossings are determined at $\pi/2$, π, 2π, ...etc.

When ζ - ω curves are used instead of σ-ω curves, the procedure is essentially the same, except that the decay time constants are determined by $1/\zeta\omega_n$ requiring each ω read from the parameter plane to be converted to ω_n. Either method gives the same results.

REFERENCES

1. Mitrovic, D., "Graphical Analysis and Synthesis of Feedback Control Systems", I - Theory and Analysis; II - Synthesis, AIEE Transactions, Pt. II, Jan. 1959.

2. Siljak, D. D., "Generalization of Mitrovic's Method", IEEE Transactions on Applications and Industry, Sept. 1964.

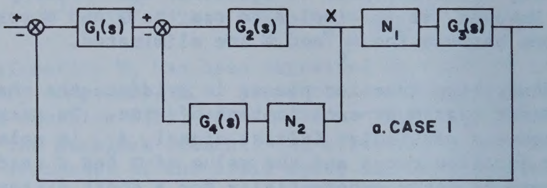
3. Siljak, D. D., "Analysis and Synthesis of Feedback Control Systems in the Parameter Plane", I - Linear Continuous Systems, III - Nonlinear Systems, IEEE Transactions on Applications and Industry, Nov. 1964.

4. Wilson, P. L., "Use of Mitrovic's Method in Predicting Limit Cycles with One and Two Gain Variable Nonlinearities", M.S. Thesis, U. S. Naval Postgraduate School, 1964.

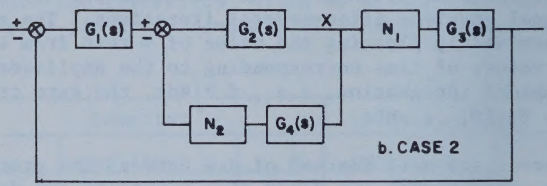
5. Yockey, H. M. and Thaler, G. J., "Analysis of Control Systems with Two Nonlinearities", Unpublished paper, U. S. Naval Postgraduate School, 1965.

6. Pellegrini, C. A., "An Investigation of Nonlinear System Performance on the Parameter Plane", M. S. Thesis, U. S. Naval Postgraduate School, 1966.

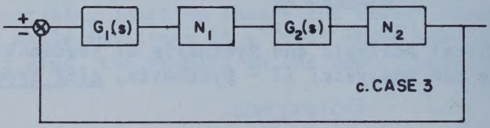
7. Siljak, D. D., "Analysis of Transient Nonlinear Oscillations in the Parameter Plane", Unpublished Paper.


+
−
$G_1(s)$
+
−
$G_2(s)$
X
N_1
$G_3(s)$
$G_4(s)$
N_2

a. CASE 1


+
−
$G_1(s)$
+
−
$G_2(s)$
X
N_1
$G_3(s)$
N_2
$G_4(s)$

b. CASE 2


+
−
$G_1(s)$
N_1
$G_2(s)$
N_2

c. CASE 3

Figure 1

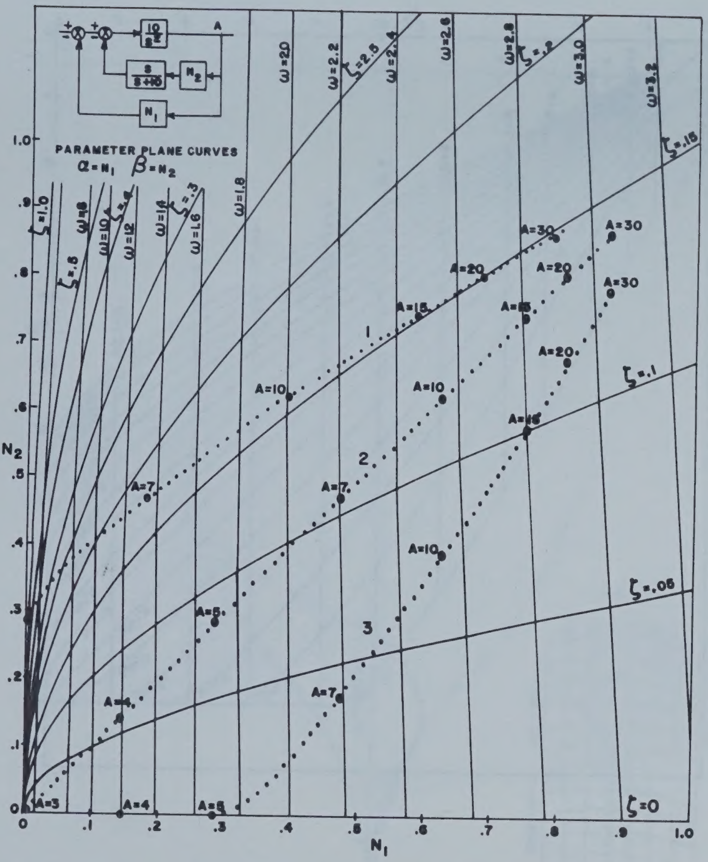


Figure 2

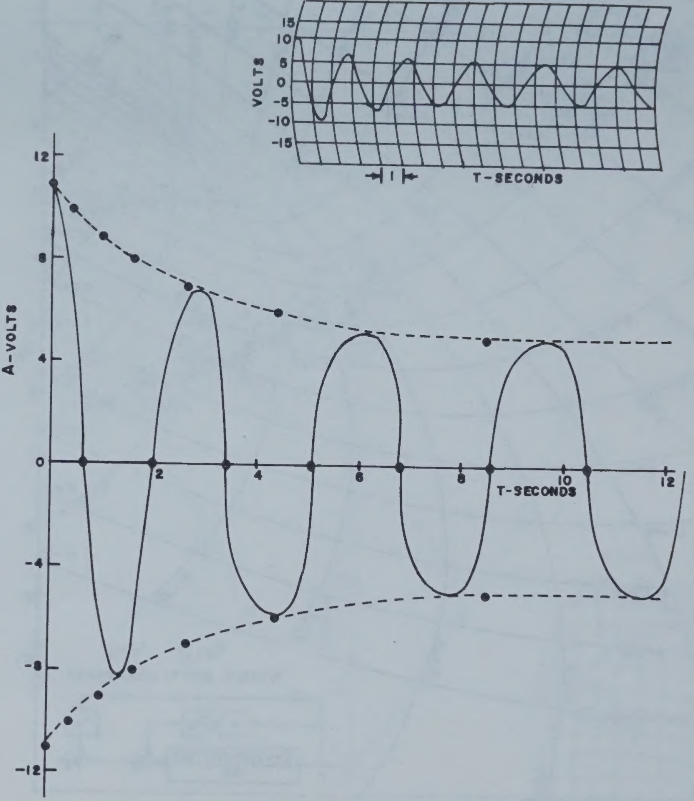
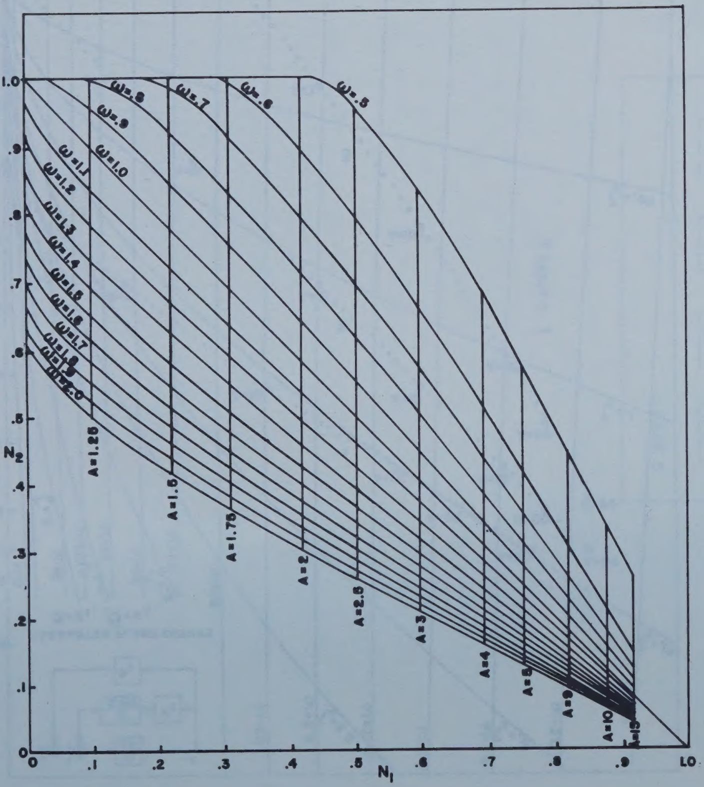
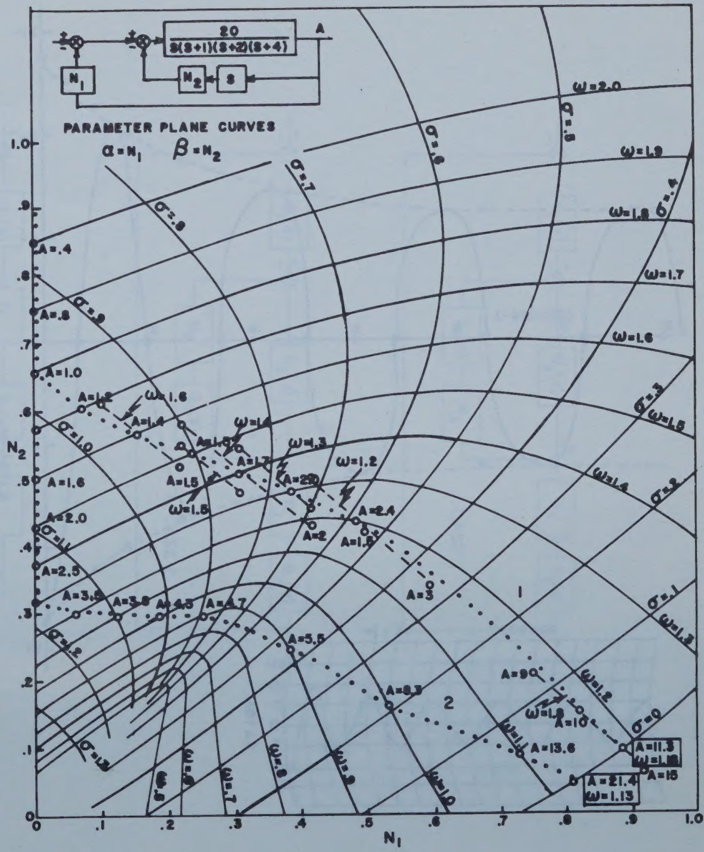


Figure 3



N₂
N₁
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
ω=.5
ω=.6
ω=.7
ω=.8
ω=.9
ω=1.0
ω=1.1
ω=1.2
ω=1.3
ω=1.4
ω=1.5
ω=1.6
ω=1.7
ω=1.8
ω=1.9
ω=2.0
A=1.25
A=1.5
A=1.75
A=2
A=2.5
A=3
A=4
A=5
A=8
A=10
A=15

Fig. 4



20
S(S+1)(S+2)(S+4)
A
N₁
N₂
S
PARAMETER PLANE CURVES
α = N₁ β = N₂
N₂
N₁
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
ω=2.0
ω=1.9
ω=1.8
ω=1.7
ω=1.6
ω=1.5
ω=1.4
ω=1.3
ω=1.2
ω=1.1
ω=1.0
ω=.9
ω=.8
ω=.7
ω=.6
ω=.5
σ=.5
σ=.6
σ=.7
σ=.8
σ=.9
σ=1.0
σ=1.1
σ=1.2
σ=1.3
σ=.4
σ=.3
σ=.2
σ=.1
σ=0
A=.4
A=.8
A=1.0
A=1.2
A=1.4
A=1.6
A=1.5
A=1.7
A=2.0
A=2.5
A=3.5
A=3.8
A=4.5
A=4.7
A=5.5
A=2
A=2.4
A=1.8
A=3
A=8.3
A=9
A=10
A=13.6
A=15
A=11.3
ω=1.18
A=21.4
ω=1.13
1
2

Fig. 5

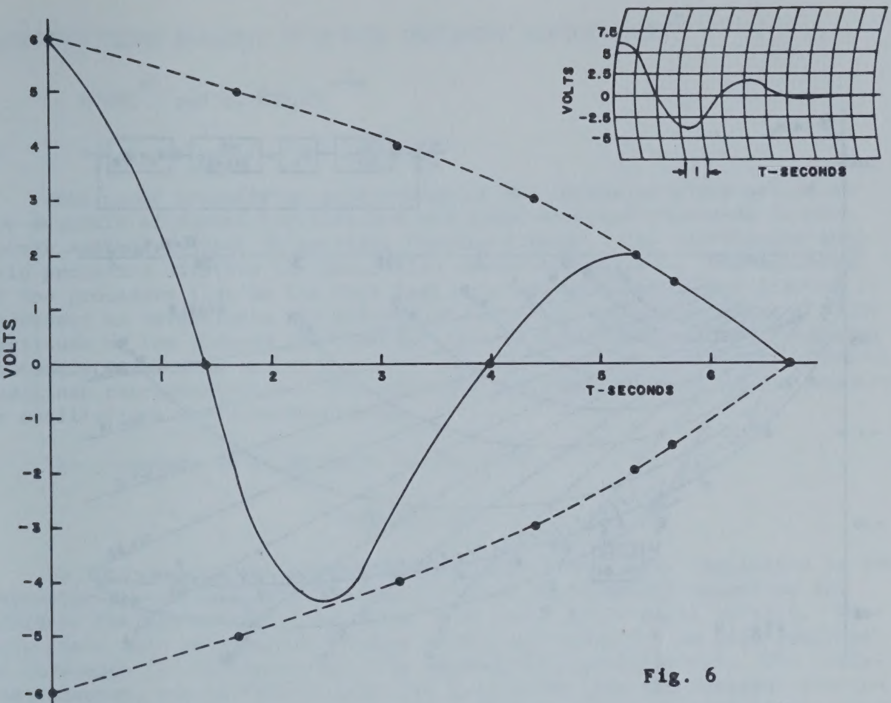

6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
VOLTS
1
2
3
4
5
6
T-SECONDS
7.5
5
2.5
0
-2.5
-5
VOLTS
1
T-SECONDS

Fig. 6

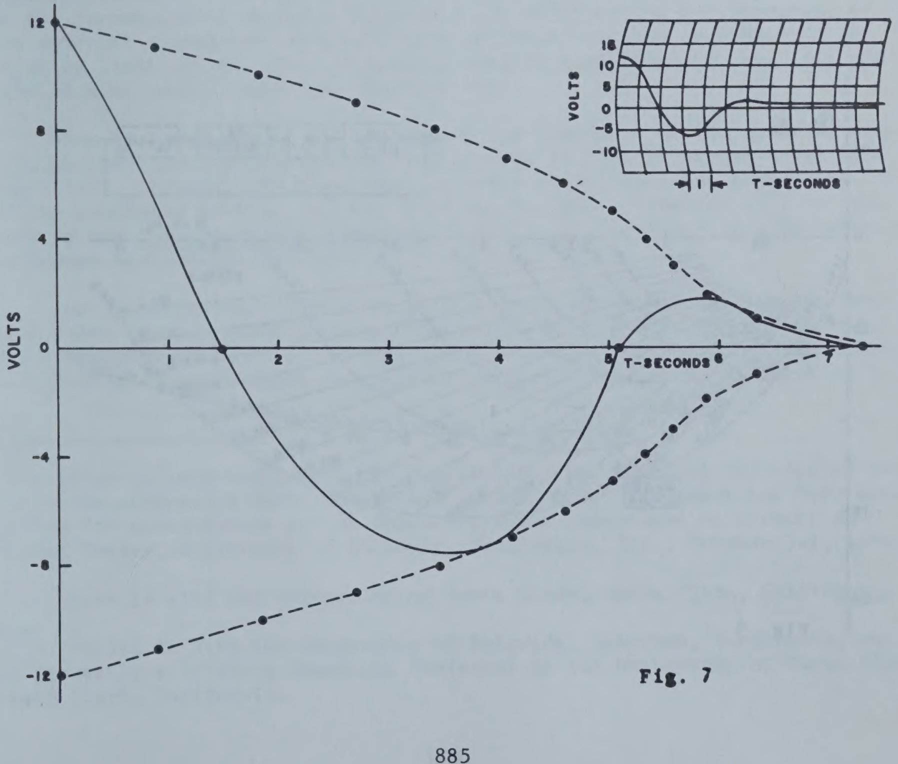
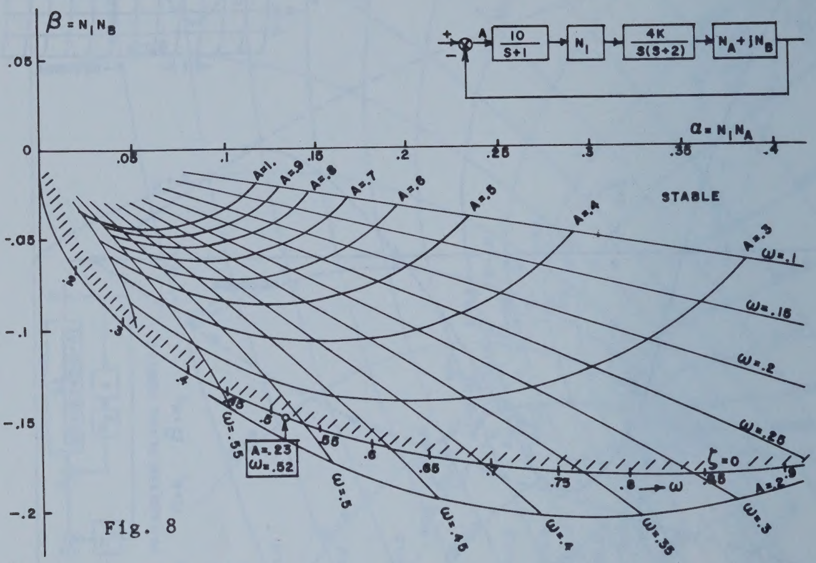
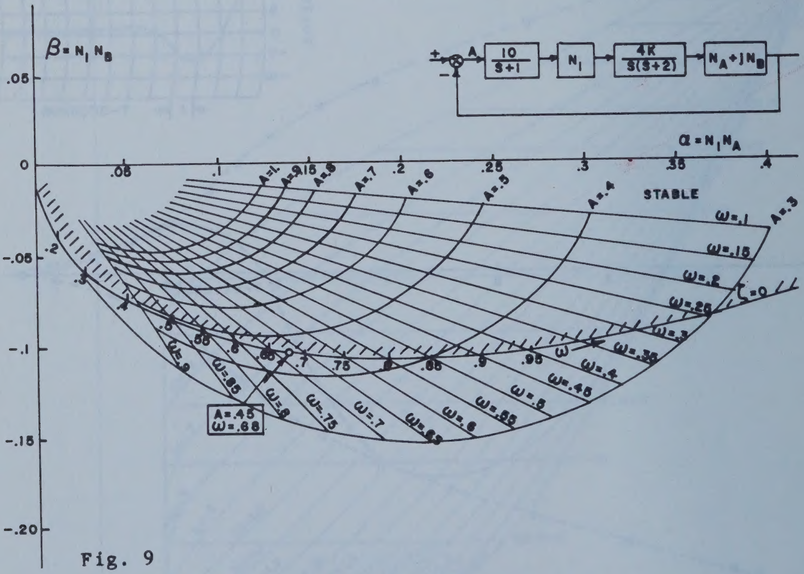

12
8
4
0
-4
-8
-12
VOLTS
1
2
3
4
5
6
7
T-SECONDS
15
10
5
0
-5
-10
VOLTS
1
T-SECONDS

Fig. 7



β = N₁N_B
α = N₁N_A
.05
0
-.05
-.1
-.15
-.2
.05
.1
.15
.2
.25
.3
.35
.4
STABLE
A=1.
A=.9
A=.8
A=.7
A=.6
A=.5
A=.4
A=.3
A=.2
ω=.1
ω=.15
ω=.2
ω=.25
ω=.3
ω=.35
ω=.4
ω=.45
ω=.5
ω=.55
.2
.3
.4
.45
.5
.55
.6
.65
.7
.75
.8 → ω
.85
.9
ζ=0
A=.23
ω=.52
+
−
A
10 / S+1
N₁
4K / S(S+2)
N_A+jN_B

Fig. 8



β = N₁N_B
α = N₁N_A
.05
0
-.05
-.1
-.15
-.20
.05
.1
.15
.2
.25
.3
.35
.4
STABLE
A=1.
A=.9
A=.8
A=.7
A=.6
A=.5
A=.4
A=.3
ω=.1
ω=.15
ω=.2
ω=.25
ω=.3
ω=.35
ω=.4
ω=.45
ω=.5
ω=.55
ω=.6
ω=.65
ω=.7
ω=.75
ω=.8
ω=.85
ω=.9
.2
.3
.4
.5
.55
.6
.65
.7
.75
.8
.85
.9
.95
ω
ζ=0
A=.45
ω=.68
+
−
A
10 / S+1
N₁
4K / S(S+2)
N_A+jN_B

Fig. 9

PARAMETER PLANE ANALYSIS OF FORCED NONLINEAR OSCILLATIONS*

J. MOORE** and D. ŠILJAK***

SUMMARY

This paper presents an application of the parameter plane method to the analysis of forced oscillations and jump-resonance phenomena in nonlinear systems subject to periodic forcing signals. The approximate analysis procedure utilizes the describing function technique. The advantage of the procedure lies in the fact that only one parameter plane diagram is necessary to investigate the effects of different forcing signals. If the amplitude of the forcing function is fixed, a modification of the proposed procedure enables an analysis of the jump-resonance phenomena for different nonlinear characteristics. The stability and sensitivity of forced nonlinear oscillations are also discussed.

The procedure is illustrated by examples.

INTRODUCTION

So far, mostly free nonlinear oscillations have been considered in the parameter plane [1-4], and the right side of differential equations describing the systems has been shown to be identically equal to zero. The only cases when an external forcing signal was present have been analyzed in reference [4] in connection with asymmetrical oscillations. The analysis, however, was performed under the assumption that the external forcing function is either constant or a slow-varying function of time with respect to the corresponding periodic solution. In other words, the frequency of the external signal has been sufficiently lower than the frequency of the existing limit cycle. This assumption greatly simplified the analysis and gave rise to useful practical applications.

When no restrictions are imposed on the frequency of the external signal, the oscillations can become complex even in case of second-order systems [5-7]. Several new phenomena can occur which could not take place in a free nonlinear system. In this section, however, attention will be focused on the jump-resonance phenomena which is of importance in a majority of nonlinear engineering problems.

The jump-resonance phenomena in high-order nonlinear systems has been long under consideration in connection with the evaluation of the closed-loop frequency response of feedback control systems containing a nonlinearity [8-20]. These methods exclusively apply the describing function tech-

---

* The research reported herein was supported by the National Aeronautics and Space Administration under Contract No. NAS2-2609. The paper has been submitted for presentation at the Fourth Allerton Conference on Circuit and System Theory, University of Illinois, Monticello, Ill., October 5-7, 1966.

** J. Moore is with the University of Santa Clara, Santa Clara, California.

*** D. Šiljak is with the University of Belgrade, Belgrade, Yugoslavia, and is presently a Visiting Associate Professor at the University of Santa Clara, Santa Clara, California.

nique. This application of the describing function to nonautonomous systems has been based upon intuition and remained suspected until recently when the more exact conditions under which such application is correct have been developed [21-24].

The methods of Levinson [8] and Prince [9] involve the definition of an equivalent gain which approximate the given nonlinear characteristic. The methods are convenient for certain single-valued limiting nonlinearities, but can hardly be adapted for more complicated nonlinear characteristics. This concept, however, has been successfully extended by Booton [25] to nonlinear systems with stochastic signals. West and others [10,15] applied the dual-input describing function to determine the condition of the jump-resonance to take place in systems with a polynomial nonlinearity. The extension to other kinds of nonlinearities is limited by the labor involved in calculating the corresponding dual-input describing function. Ogata and Hopkins' analytic techniques [11,13] are limited to relay characteristics, or, perhaps, saturating nonlinearities, since the output wave shape of the nonlinearity should be assumed. Considerable work is required to make a proper assumption in cases other than the ideal relay characteristic. Stein and Thaler [12] proposed a trial-and-error procedure to calculate the closed-loop frequency response by using the Nichols chart. A similar concept has been used with some modifications in reference [17] by McAllister.

The idea of finding conditions under which the jump-resonance occurs and then designing the linear part of the system to avoid that as proposed by West and others [10] has been successfully applied along with the common describing function in reference [15,16,18]. The methods, however, cannot give an insight into the frequency response of the closed-loop systems. Moreover, the procedures are extremely cumbersome when applied to the simplest multi-valued nonlinearities or nonlinearities with the frequency dependent describing functions.

The frequency response can be analyzed by a different approach proposed by Gibson [19], which has some apparent advantages over the previously presented procedures. It enables information to be readily obtained about the effects on the frequency response of varying the nonlinear parameters. In addition, it does not require the analytical expression of the describing function and can thus be applied to experimentally obtained data. The procedure, however, has to be repeated each time the amplitude of the forcing periodic signal is changed. This can be circumvented by the technique of Popov and Palitov [14] at the expense that the nonlinearity cannot be varied without recalculating all the necessary curves.

In this paper, the separate ideas of the references mentioned in the preceding paragraph will be employed using the parameter plane concept. The proposed analysis is directed towards the consideration of jump phenomena in high-order nonlinear systems. In addition, the stability and sensitivity of the periodic oscillations is briefly discussed on the basis of reference [3]. The method can be extended directly to the analysis of subharmonic resonance by employing the dual-input describing function as described by West [15].

Periodic Solutions. Jump-Resonance.

Consider again the nonlinear differential equation

$$B(s)x + C(s)\,F(x,sx) = H(s)f \qquad s \equiv \frac{d}{dt} \tag{1}$$

where, at the right hand side, the forcing function $f = f(t)$ is

$$f = A_f \sin(\Omega_f t - \varphi) \tag{2}$$

and B(s), C(s), and H(s) are polynomials in s. C(s) and H(s) are polynomials with degrees less than that of the polynomial B(s). The function $F(x,sx)$ represents the nonlinearity. The basic problem is to determine the conditions under which equation (1) has a periodic solution $x = x(t)$ sufficiently close to

$$x = A \sin(\Omega_f t) \tag{3}$$

where the frequency Ω_f is the known frequency of the forcing function $f(t)$ in (2); and then to evaluate the unknown values of the amplitude A and the phase-shift φ.

By a suitable transformation, the above problem can be reduced to that of free symmetrical oscillations considered in references [1-3]. To do this, note that the function f can be related to x by deriving

$$f = A_f \cos\varphi \sin(\Omega_f t) - A_f \sin\varphi \cos(\Omega_f t) \quad ,$$

from (3) and noting that

$$sx = A\,\Omega_f \cos(\Omega_f t),$$

one finally has

$$f = \frac{A_f}{A}\left(\cos\varphi - \frac{\sin\varphi}{\Omega_f} s\right)x.$$

By substituting this expression of f into equation (1), one obtains

$$\left[B(s) - H(s)\frac{A_f}{A}\left(\cos\varphi - \frac{\sin\varphi}{\Omega_f} s\right)\right]x + C(s)\,F(x,sx) = 0 \quad .$$

Therefore, the nonhomogeneous differential equation (1) is reduced to a homogeneous one by knowing the forcing function $f = f(t)$ and by assuming the form of the solution $x = x(t)$. To determine the amplitude A and the phase-shift φ of the solution $x(t)$, the methods for the analysis of symmetrical self-excited oscillations as outlined in [1-3] can be used with minor modifications.

The nonlinearity $F(x,sx)$ can be harmonically linearized as

$$F(x,sx) = N_1 x + \frac{N_2}{\Omega_f} sx$$

where the coefficients $N_1 = N_1(A,\Omega_f)$ and $N_2 = N_2(A,\Omega_f)$ are the describing function coefficients,

$$N_1 = \frac{1}{\pi A} \int_0^{2\pi} F(A \sin\phi,\, A\,\Omega_f \cos\phi) \sin\phi \, d\phi$$

$$N_2 = \frac{1}{\pi A} \int_0^{2\pi} F(A \sin \phi, A \Omega_f \cos \phi) \cos \phi \, d\phi$$

and $\phi = \Omega_f t$. The standard tables of reference [16] may be used for calculating N_1 and N_2 .

The linearized differential equation has the form

$$[B(s) - H(s) \frac{A_f}{A} (\cos \varphi - \frac{\sin \varphi}{\Omega_f} s) + C(s)(N_1 + \frac{N_2}{\Omega_f} s)]x = 0 \qquad (4)$$

and the corresponding characteristic equation is

$$B(s) - H(s) \frac{A_f}{A} (\cos \varphi - \frac{\sin \varphi}{\Omega_f} s) + C(s)(N_1 + \frac{N_2}{\Omega_f} s) = 0 \qquad (5)$$

The periodic solution $x = a \sin (\Omega_f t)$ with the frequency Ω_f can be determined from (5) by substituting $s = j\Omega_f$ and using the condition that the summation of reals and imaginaries must go to zero independently. Thus, by certain simple algebraic manipulations, one obtains

$$\begin{aligned} B_1 A - H_1 A_f \cos \varphi - H_2 A_f \sin \varphi + C_1 A N_1 - C_2 A N_2 &= 0 \\ B_2 A + H_1 A_f \sin \varphi - H_2 A_f \cos \varphi + C_1 A N_2 + C_2 A N_1 &= 0 \end{aligned} \qquad (6)$$

where $B_1 = B_1(\Omega_f)$,$B_2=B_2(\Omega_f)$, etc., are given as

$$B_1 = \sum_{k=o}^{n} b_k X_k \, , \qquad B_2 = \sum_{k=o}^{n} b_k Y_k \, , \quad \text{etc.}$$

The functions X_k and Y_k are calculated using the recurrence relationships [see reference 26]

$$X_{k+1} + \Omega^2 X_{k-1} = 0$$

$$Y_{k+1} + \Omega^2 Y_{k-1} = 0$$

where $X_o \equiv 1$, $X_1 \equiv 0$, $Y_o \equiv 0$, $Y_1 \equiv \Omega$. Due to the above recurrence relationships, the $\zeta = 0$ curve can be readily plotted using a general digital computer program for n-th order characteristic polynomials.

By denoting

$$\begin{aligned} \alpha &= A_f \cos \varphi \\ \beta &= A_f \sin \varphi \end{aligned} \qquad (7)$$

equations (6) may represent two equations in two unknowns, α and β which can be solved for α and β as

$$\alpha = - A \frac{H_1(B_1 + C_1N_1 - C_2N_2) + H_2(B_2 + C_1N_2 + C_2N_1)}{H_1^2 + H_2^2} \qquad (8)$$

$$\beta = A \frac{H_1(B_2 + C_1N_2 + C_2N_1) - H_2(B_1 + C_1N_1 - C_2N_2)}{H_1^2 + H_2^2}$$

In the $\alpha\beta$-plane, equations (8) may be interpreted as equations of the $\zeta = 0$ curve as defined in reference [2]. The solution procedure starts with the plotting of a family of $\zeta = 0$ curves for different values of the frequency Ω_f appearing in the coefficients, B_1, B_2, C_1, C_2, H_1, N_1 and N_2 of (8). The unknown amplitude A, which enters both explicitly and as an argument of N_1 and N_2 in (8), is interpolated along the $\zeta = 0$ curves. The loci of the point $M(\alpha = A_f \cos \varphi \ ; \ \beta = A_f \sin \varphi)$ are concentric circles with a radius A_f and the phase-shift φ interpolated along the circles. Once the ζ and M-point loci are plotted in the $\alpha\beta$-plane, the amplitude A, frequency Ω_f and the phase-shift φ of the possible periodic solution $x = A \sin (\Omega_f t)$ are determined graphically from the obtained plot in a straightforward manner as shown in the following example*.

Consider a control system with the block diagram shown on Fig. 1. The related differential equation has the form

$$s(s+0.5)(s+2)(s+10)x + 10(s+1)F(x) = s(s+0.5)(s+2)(s+10)f \qquad (9)$$

where the variable $x = x(t)$ represents the error signal $e(t)$, and the forcing function $f = f(t)$ is the input $r(t)$.

For the sinusiodal input function $r(t) = A_f \sin(\Omega_f t - \varphi)$ the characteristic equation of the corresponding linearized system is

$$s(s+0.5)(s+2)(s+10)[A - A_f(\cos \varphi - \frac{\sin \varphi}{\Omega_f} s)] + 10(s+1) AN_1(A) = 0 \quad (10)$$

where the solution $e(t)$ of equation (9) is assumed as $e(t) = A \sin (\Omega_f t)$.

After the substitution

$$\alpha = A_f \cos \varphi \quad , \qquad \beta = A_f \sin \varphi$$

---

*This example has been treated analytically in reference [13] whereby, for every set of values (A_f, Ω_f), at least one transcedental equation has to be solved.

and $s = j\Omega_f$ into equation (10), one obtains the equation (8) of the $\zeta = 0$ curve as

$$\alpha = \alpha(A, \Omega_f)$$

$$\beta = \beta(A, \Omega_f)$$

For the specific equation (10), equation (7) of the M-point loci and equations (8) of the $\zeta = 0$ curve are plotted on Fig. 2 for different values of the amplitude A_f and the frequency Ω_f, respectively, as family parameters.

The diagram of Fig. 2 is interpreted in the usual manner. For example, if the amplitude and frequency of the forcing signal $f(t)$ are $A_f = 5$ and $\Omega_f = 2$, then the related point is M_o for which the periodic solution $x(t)$ has the amplitude $A = 9.3$ and the phase-shift $\varphi \approx -19^\circ$.

It is of particular significance to note that if the frequency Ω_f is increased up to $\Omega_f = 1.4$ rad/sec and keeping $A_f = 5$, the same M locus intersects the $\zeta = 0$ curve at three points, M_1, M_2, and M_3, which corresponds to three different periodic solutions with the same frequency $\Omega_f = 1.4$ rad/sec. This is due to the so-called jump resonance which can be better understood by plotting the frequency response characteristics from the diagram of Fig. 2.

In Fig. 3, the closed-loop frequency response relating the input and output of the system under investigation is plotted by an analog computer for the input amplitude $A_f = 5$. The jump resonance is indicated in between the frequencies $\Omega_f = 1$ and $\Omega_f = 2$. When the frequency Ω_f of the input signal is gradually increased, the first part 0ab of the gain curve $\frac{C}{R}(j\Omega_f)$ is obtained on Fig. 3. When the frequency Ω_f is increased at the point b, the system output changes abruptly reaching the point c. With a further increase in the input frequency, the continuous part cd is obtained. If the procedure is reversed and the input frequency is decreased gradually from a higher value, the part dcb' of the gain curve is plotted. When the point b' is reached, a decrease in Ω_f causes the output amplitude to drop discontinuously to the point a and then follows the part a0 as the frequency Ω_f decreases to zero. The corresponding phase characteristic follows the discontinuous jumps in the amplitude as shown in the same Fig. 3.

The three points 1, 2, 3 on the gain characteristic of Fig. 3 correspond to the points M_1, M_2, M_3 in the diagram of Fig. 2. The lower point 1 and the upper point 3 are related to the stable periodic solution, while the middle point 2 is unstable and is not observed in the experiments. The jump resonance obtained by the computer, checks the results available on the diagram of Fig. 2. For example, the $\zeta = 0$ curve for $\Omega_f = 1.3$ is tangent to the M locus plotted for $A_f = 5$. This corresponds to the points a and b'. By interpolating between the $\zeta = 0$ curves plotted for $\Omega_f = 1.4$ and $\Omega_f = 1.5$, the frequency of the jump bc can be evaluated. (Note that the diagram of Fig. 2 relates the input $r(t)$ and the error signal $e(t)$, while the plot of Fig. 3 relates $r(t)$ and the output $c(t)$. This is of no essential importance since a simple relationship $e = r-c$ holds.)

Another observation of the jump resonance can be made if the input frequency is held constant but the input amplitude is varied so that A_f varies while Ω_f is constant. By using again the diagram of Fig. 2, it is not difficult to show that if Ω_f is chosen to be 1.4 and A_f is varied, a diagram of Fig. 4 can be calculated from that of Fig. 2. Once again there is a range of values of A_f for which three values of A are possible. For $A_f = 5$, there are three values, A_1, A_2, A_3, of the amplitude A which correspond to the three points M_1, M_2, M_3 of Fig. 2. The lower A_1 and upper A_3 correspond to stable solutions, while A_2 is unstable and cannot be obtained experimentally. The corresponding phase diagram is shown in Fig. 5.

It is of interest to note all diagrams of Figs. 3,4 and 5 are obtained for specific values of either the amplitude A_f (Figs. 3, $A_f = 5$) or the frequency Ω_f (Figs. 4 and 5, $\Omega_f = 1.4$). If these values are changed, all the diagrams have to be replotted again since they cannot be normalized with respect to the input as is possible in linear systems. The effects on (A, φ) of changing $(A_f \Omega_f)$, however, can be studied directly from the diagram of Fig. 2. Furthermore, the presented procedure can be applied equivalently to single-valued and common multi-valued nonlinearities as well as to nonlinearities with the frequency-dependent describing functions. In certain cases, the procedure can be extended to the analysis of the jump phenomena in nonlinear systems with two nonlinearities provided the applicability conditions of the describing function are satisfied.

Stability and Sensitivity of the Periodic Solutions.

In general, the jump phenomena can be more complex than that examined in the previous section. It may contain more than one curl and, thus, more than three periodic solutions are possible. After these solutions are determined, the stability problem arises to separate the stable from the unstable solutions.

The unstable solutions are related to the case when an increase in the input amplitude A_f results in a decrease of the amplitude A of the corresponding periodic solution. By examining Fig. 4, one may conclude that the unstable solution amplitudes are located along the part of the curve for which the slope is negative. The stability can now be checked from the parameter plane diagram. So, for an increase in the amplitude A along the curve $\Omega_f = 1.4$ at the point M_1, there is an increase in the amplitude A_f, and the related solution is stable. The same reasoning reveals that the point M_2 is unstable, but M_3 is again stable.

The stability of periodic solutions can be checked analytically by determining the sign of the derivative $\partial A/\partial A_f$. This derivative is calculated from equations 6 for the case when the frequency Ω_f is considered constant and the amplitude A and phase φ are assumed to be functions of the amplitude A_f . This is quite similar to the procedure of the sensitivity analysis of self-excited nonlinear oscillations presented in reference [3]. Thus, it is left to the reader to derive the derivatives $\partial A/\partial A_f$ and $\partial\varphi/A_f$ from equations 6; then, to apply the obtained expressions to the system investigated in the preceding section.

Likewise, the sensitivity analysis is related to the stability problem and the concept of sensitivity analysis of small parameter variations in self-excited oscillations developed in reference [3] can be extended to the forced oscillations. The coefficients of the polynomials B(s), C(s), and H(s) in the characteristic equation

$$B(s) - H(s)\frac{A_f}{A}\left(\cos\varphi - \frac{\sin\varphi}{\Omega_f}\right) + C(s)\left(N_1 + \frac{N_2}{\Omega_f}s\right) = 0 \qquad (5)$$

can be considered as functions of μ linear parameters q_i, $(i=1,2,\ldots,\mu)$. The describing function $N = N_1 + jN_2$ can be considered as a function of A, Ω_f, and ν nonlinear parameters p_j, $(j = 1,2,\ldots,\nu)$. Then the sensitivities

$$S^A_{q_i} \overset{\Delta}{=} \frac{\partial \ell n\, A}{\partial \ell n\, q_i}, \qquad S^{\varphi}_{q_i} \overset{\Delta}{=} \frac{\partial \ell n\, \varphi}{\partial \ell n\, q_j}$$

$$S^A_{p_j} \overset{\Delta}{=} \frac{\partial \ell n\, A}{\partial \ell n\, p_j}, \qquad S^{\varphi}_{p_j} \overset{\Delta}{=} \frac{\partial \ell n\, \varphi}{\partial \ell n\, p_j}$$

can be calculated by differentiating equations (6) with respect to either q_i or p_j . The obtained sensitivity values indicates how the steady-state values of the amplitude A and phase φ are affected by the small parameter variations.

It is of interest to note that the possibility of the jump resonance to take place can be indicated by the condition that the derivative $\partial A/\partial A_f = \infty$; i.e., $\partial A_f/\partial A = 0$. This condition has been examined entirely in reference [16,17] whereby a graphical procedure has been proposed to check the condition being satisfied. The procedure of reference [17] is convenient for adjusting the linear part of the system so that jump resonance is avoided either entirely or for some range of input amplitudes.

Variation of Nonlinear Characteristic.

In the case of single-valued nonlinear characteristics when the nonlinear differential equation has the form

$$B(s)x + C(s)\,F(x) = H(s)f$$

and the forcing function $f = f(t)$ is

$$f(t) = A_f \sin(\Omega_f t - \varphi) \tag{2}$$

the analysis of the periodic solution $x = x(t)$

$$x(t) = A \sin(\Omega_f t) \tag{3}$$

can be performed by a different approach proposed by Gibson [19]. This approach is convenient for the analysis of the effects on jump resonance of varying the nonlinear characteristic; i.e., the nonlinear parameters. Moreover, the approach can be advantageous in cases when the nonlinear single-valued characteristic is found experimentally and the describing function is determined graphically. Of course, it may be applied to situations when the describing function of the nonlinearity is plotted experimentally or by using computers. The approach will be generalized by the parameter plane concept.

To describe the approach, note that the same harmonic linearization presented in the preceding section if applied to equation (10) yields

$$[B(s) + H(s) \frac{A_f}{A} (\cos \varphi - \frac{\sin \varphi}{\Omega_f} s + C(s)N_1]x = 0 \tag{11}$$

where $N_1 = N_1(A)$. Equation (11) can be obtained from equation (4) by assuming $N_2 \equiv 0$, which is true for a single-valued nonlinearity $F(x)$. The corresponding characteristic equation can be written as

$$B(s)A + C(s)\, AN_1 + H(s)\, A_f(\cos \varphi - \frac{\sin \varphi}{\Omega_f} s) = 0 \quad . \tag{12}$$

After substituting $s = j\Omega_f$ into (12), one obtains two equations in two unknowns, α and β . Thus,

$$B_1 \alpha + C_1 \alpha\beta + H_1 A_f \cos \varphi + H_2 A_f \sin \varphi = 0$$

$$B_2 \alpha + C_2 \alpha\beta - H_1 A_f \sin \varphi + H_2 A_f \cos \varphi = 0 \tag{13}$$

where

$\alpha = A$

$\beta = N_1(A)$

Equation (13) can be solved for α and β to obtain

$$\frac{\alpha}{A_f} = \frac{(C_2H_1 - C_1H_2)\cos\varphi + (C_2H_2 + C_1H_1)\sin\varphi}{B_2C_1 - B_1C_2}$$

$$\beta = \frac{(B_1H_2 - B_2H_1)\cos\varphi - (B_2H_2 + B_1H_1)\sin\varphi}{(C_2H_1 - C_1H_2)\cos\varphi + (C_2H_2 + C_1H_1)\sin\varphi} \qquad (14)$$

Now, for a given value of the forcing input frequency Ω_f, the coefficients B_1, B_2, C_1, C_2, H_1, and H_2, which are functions of Ω_f, only, are evaluated numerically. For different values of the phase shift φ, the corresponding values α and β are calculated from (14) Thus a locus with constant frequency Ω_f can be plotted in the $\alpha\beta$ plane where the scale factor along the α axis is given by A_f. On the other hand, the M locus is simply the describing function curve itself. Any change in the nonlinear parameters affects only the M locus while any change in the A_f input amplitude simply effects the scale along the α axis of the $\alpha\beta$ plane for the M locus only.

To illustrate the analysis procedure, consider the same example of the previous section. Applying the outlined procedure, one obtains the parameter plane diagram shown in Fig. 6. The diagram is plotted for $A_f = 5$ and, therefore, it can be compared with the computer simulation diagram of Fig. 3 to indicate the accuracy. The points 1,2,3 of the intersections of the M locus with the $\Omega_f = 1.4$ corresponds to the points 1,2,3 of Fig. 3. The corresponding values of the phase shift φ are read on the Ω_f curves.

Any change in the amplitude A_f can be interpreted merely as a shift of the M locus relative to the Ω_f-constant curves of Fig. 6. Thus, a slight modification of the M locus should be made each time the amplitude A_f is changed.

Conclusions.

It has been shown that forced nonlinear oscillations, and in particular the jump-resonance phenomena may be conveniently considered using the parameter plane concept. Either a parameter plane may be plotted for a specified nonlinearity and the effects of varying the forcing signal amplitude and frequency may be considered, or a parameter plane may be plotted such that the effects of varying the nonlinear characteristic may be investigated at the expense of ease of interpretation for input amplitude changes. The advantage of using the parameter plane approach is that the required design information is given conveniently on the one diagram.

REFERENCES

1. Šiljak, D.D., "Analysis and Synthesis of Feedback Control Systems in the Parameter Plane", Pt. I - Linear Continuous Systems; Pt II - Sampled-Data Systems; Pt. III - Nonlinear Systems, IEEE Transactions on Applications and Industry, Nov., 1964, pp 458-473.

2. Šiljak, D.D.,"Generalization of the Parameter Plane Method", IEEE Transactions on Automatic Control", vol. AC-11, Jan., 1966, pp 63-70.

3. Šiljak, D.D., and M. R. Stojić', "Sensitivity Analysis of Self-Excited Nonlinear Oscillations", IEEE Transactions on Automatic Control, vol. AC-10, Oct., 1965, pp 413-420.

4. Šiljak, D.D., "Analysis of Asymmetrical Nonlinear Oscillations in the Parameter Plane", IEEE Transactions on Automatic Control, vol. AC-11, April, 1966, pp 239-247.

5. Bogoliubov, N.N., and U.A. Mitropolsky, "Asymptotical Methods in the Theory of Nonlinear Oscillations", (in Russian), Fizmatgiz, 1963. (English Trans., Delhi, India: Hindustan Pub. Co., USA Dist.: Gordon and Breach Sci. Pub., New York).

6. Cunningham, W.J., "Introduction to Nonlinear Analysis", McGraw-Hill Book Co., Inc., New York, 1958.

7. Hayashi, Ch., "Oscillations in Physical Systems", McGraw-Hill Book Co., Inc., New York, 1964.

8. Levinson, E., "Some Saturation Phenomena in Servomechanisms with Emphasis on the Tachometer Stabilized System", AIEE Trans., vol. 72, Pt II (Applications and Industry), March 1953, pp 1-9.

9. Prince, L.T., Jr., "A Generalized Method for Finding the Closed-loop Frequency Response of Nonlinear Systems", AIEE Trans., vol. 73, Pt II (Applications and Industry), Sept., 1954, pp 217-224.

10. West, J.C., J.L. Douce, and R.K.Livesley, "The Dual-Input Describing Function and Its Use in the Analysis of Nonlinear Feedback Systems", IEEE Proc., vol. 103, B, 1956, pp 463-474.

11. Ogata, K., "An Analytic Method for Finding the Closed-loop Frequency Response of Nonlinear Feedback-control Systems", AIEE Trans., vol. 76, Pt II (Applications and Industry), Nov., 1957, pp 277-285.

12. Stein, W.A., and G. J. Thaler, "Obtaining the Frequency Response Characteristic of a Nonlinear Servomechanism for an Amplitude- and Frequency-sensitive Describi-g Function", AIEE Trans., vol. 77, Pt II (Applications and Industry), May, 1958, pp 91-96.

13. Hopkin, A.M., and K. Ogata, "An Analytic Frequency Response Solution for a Higher Order Servomechanism with a Nonlinear Control Element", J. of Basic Engr., ASME Trans., Series D. Vol. 81, March 1959, pp 41-45.

14. Popov, E.P., and Palitov, "Approximate Methods for Analyzing Nonlinear Automatic Systems:, (in Russian), Fizmatgiz, Moscow, 1960. (English trans. Foreign Tech. Div., AFSC, Wright-Patterson Air Force Base, Ohio, Rept. FTD-TT-62-910.)

15. West, J.C., "Analytical Techniques for Nonlinear Control Systems", D. Van Nostrand Co., Inc., Princeton, New Jersey, 1960.

16. Gille, J.C., and P. Decaulne, "On the Forcing Oscillations of Nonlinear Servosystems" (in French), Automatic and Remote Control, Butterworth London, vol. 1, 1961, pp 205-210.

17. McAllister, A.S., "A Graphical Method for Finding the Frequency Response of Nonlinear Closed-Loop Systems, AIEE Trans., vol. 80, Pt II (Applications and Industry), Nov., 1961, pp 268-277.

18. Hatanka, H., "The Frequency Response and Jump-resonance Phenomena of Nonlinear Feedback Control Systems:, J. of Basic Engr., ASME Trans., Series D, Vol. 85, June, 1963, pp 236-242.

19. Gibson, J.E., "Nonlinear Automatic Control", McGraw-Hill Book Co., Inc. New York, 1963.

20. Fukuma, A., and M. Matsubara, "Jump-Resonance Criteria of Nonlinear Control Systems", (to be published in IEEE Trans. on Automatic Control).

21. Sagirow, P., "On the Problem of Error Estimation in the Harmonic Balance" (in German), Z. augen. Math. und Mech., Nr. 10-11, 1960.

22. Sagirow, P., "On the Periodic Solutions of Nonlinear Differential Equations and on the relation between the Amplitude and Frequency" (in German), Z. augen. Math. und Mech., Nr. 3, 1961.

23. Garber, E.D., "An Estimate of the Error of the Harmonic Balance Method" (in Russian) Automatika i telemekhanika, t. 24, No. 4, April, 1963, pp. 482-492.

24. Garber, E.D., and E.N. Rozenvasser, "The Investigation of Periodic Regimes of Nonlinear Systems on the Basis of the Filter Hypothesis" (in Russian), t. 26, No. 2, Feb., 1965, pp 277-287.

25. Booton, R.C., Jr., "The Analysis of Nonlinear Control Systems with Random Inputs", Proceedings of the Symposium on Nonlinear Circuit Analysis, New York, Polytechnic Institute of Brooklyn, 1953, pp 369-391.

26. Šiljak, D.D., "A Note on the Generalized Nyquist Criterion", IEEE Transactions on Automatic Control, vol. AC-11, No. 2, April, 1966, pp 317.

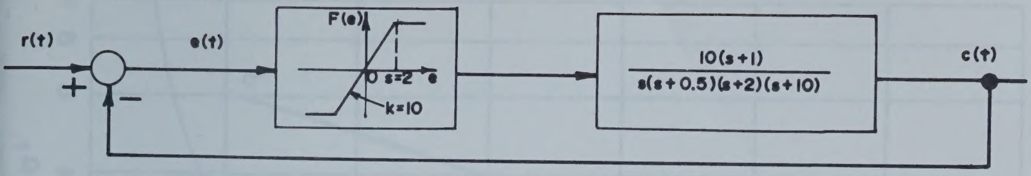

r(t)
+
−
e(t)
F(e)
0
s=2
e
k=10
10(s+1)
s(s+0.5)(s+2)(s+10)
c(t)

FIG. 1

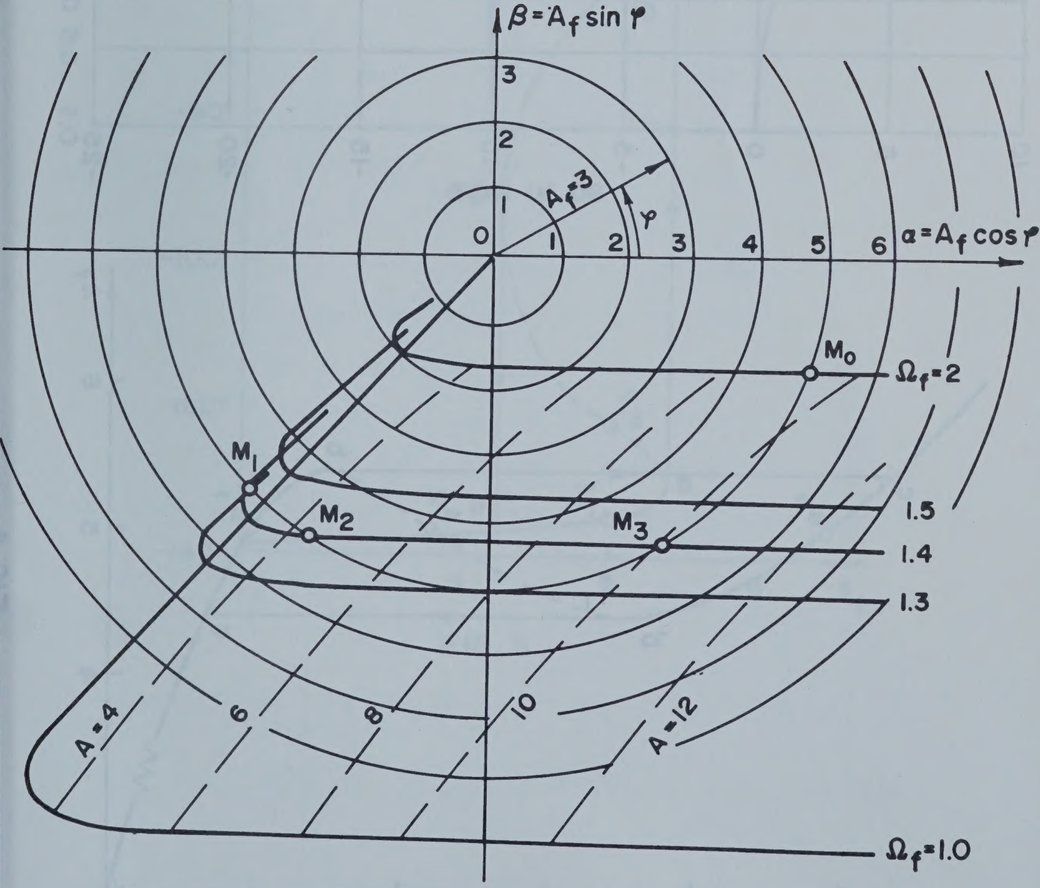

$\beta = A_f \sin \varphi$
$\alpha = A_f \cos \varphi$
$A_f = 3$
φ
0
1
2
3
4
5
6
M_0
M_1
M_2
M_3
$\Omega_f = 2$
1.5
1.4
1.3
$\Omega_f = 1.0$
A=4
6
8
10
A=12

FIG. 2

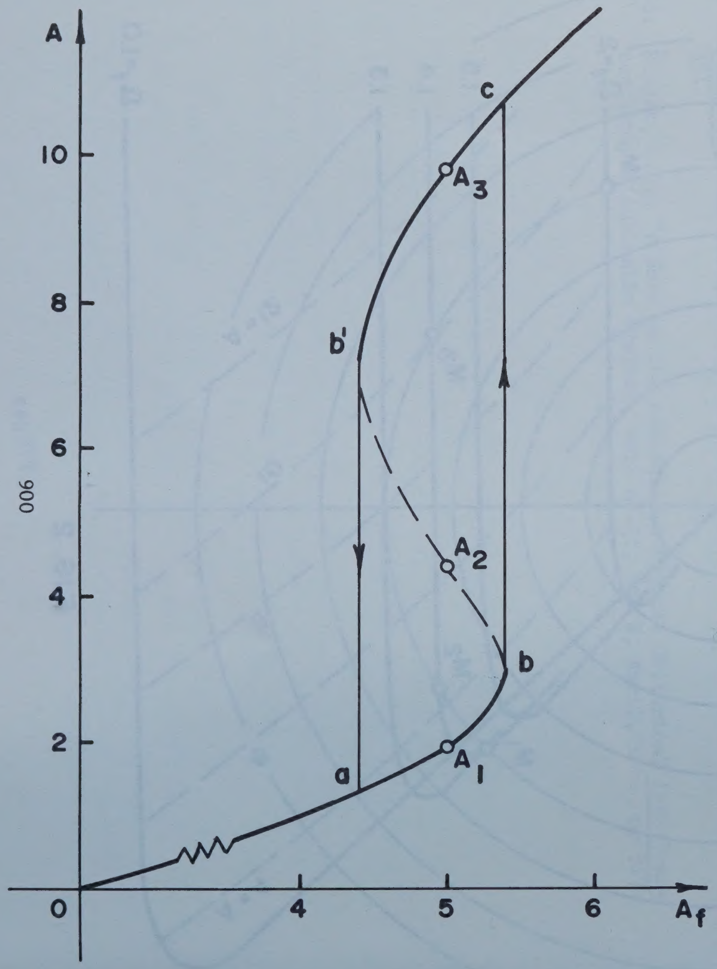

A
10
8
6
4
2
0
4
5
6
A_f
c
A_3
b'
A_2
b
A_1
a

FIG. 4

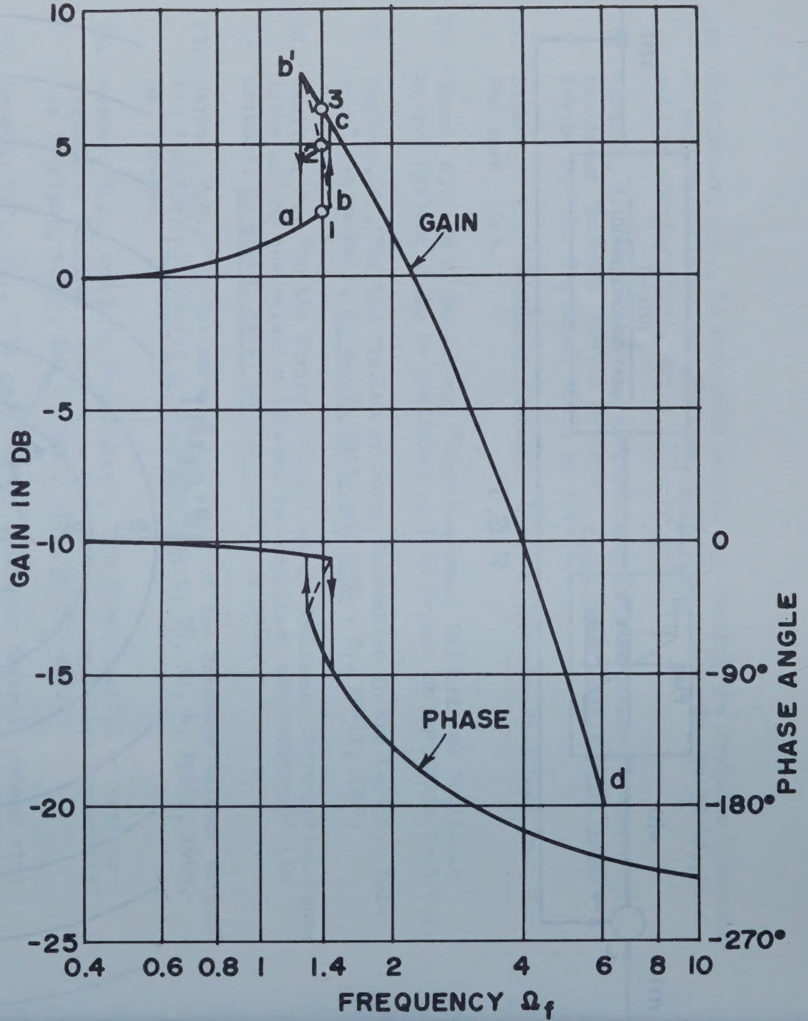

GAIN IN DB
10
5
0
-5
-10
-15
-20
-25
0.4
0.6
0.8
1
1.4
2
4
6
8
10
FREQUENCY Ω_f
PHASE ANGLE
0
-90°
-180°
-270°
b'
3
c
2
b
a
1
GAIN
PHASE
d

FIG 3

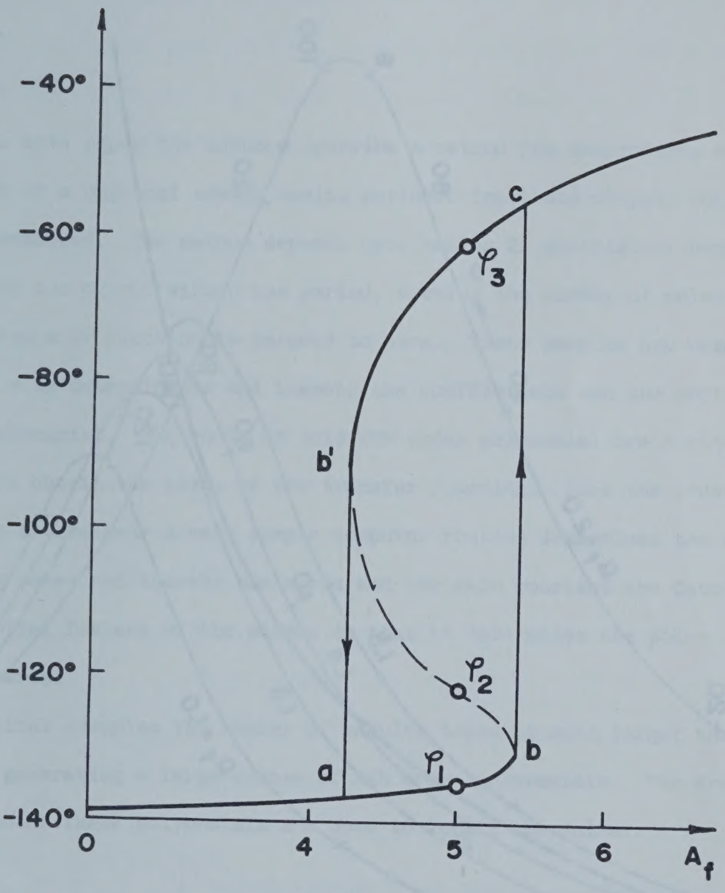


FIG. 5

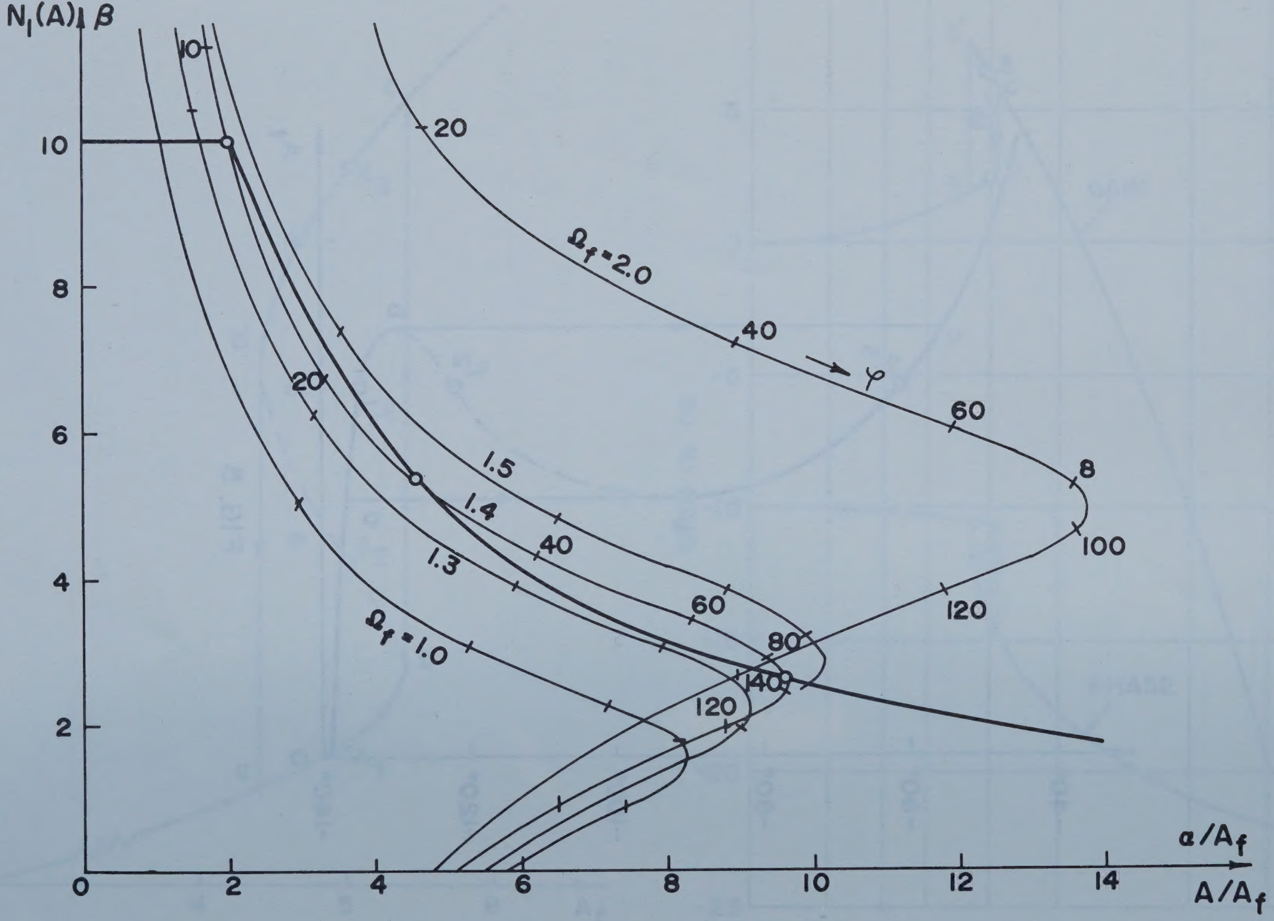

$N_1(A)$, β
10
8
6
4
2
0
2
4
6
8
10
12
14
a/A_f
A/A_f
$\Omega_f = 2.0$
1.5
1.4
1.3
$\Omega_f = 1.0$
10
20
20
40
40
60
60
80
100
120
120
140
8
φ

FIG.6

IDENTIFICATION OF A LINEAR SYSTEM FROM DISCRETE VALUES OF INPUT-OUTPUT DATA

BY

C.N. WEYGANDT[+] AND N.N. PURI*

Summary:

In this paper the authors describe a method for determining the transfer function of a physical system having periodic input and output, both of which can be measured. The method depends upon taking 2n equidistant samples of the input and the output within one period, n being the number of poles the system transfer function is assumed to have. These samples are used to compute (n + 1) determinants and thereby the coefficients and the roots of an nth order polynomial. The roots of this nth order polynomial are further processed to obtain the poles of the transfer function. Once the location of poles is determined, a very simple computer routine determines the residues of these poles and thereby the zeros and the gain constant are determined. An essential feature of the scheme is that it determines the poles independent of the zeros.

In practical examples the number of samples taken is much larger than 2n, thereby generating a large number of nth order polynomials. The average of the roots of these polynomials are used to locate the poles.

[+] Professor, Moore School of Electrical Engineering, University of Penna., Philadelphia, Pennsylvania.

* Consultant, General Electric, Re-Entry Systems Division, 3198 Chestnut Street, Philadelphia, Pennsylvania

1. INTRODUCTION

In many branches of science there are situations where the input and output of a linear system are known as recorded curves or as digital samples obtained at uniform time intervals. This paper presents a method of identifying such a system in terms of the poles, zeros and gain of the transfer function between the input and the output. Thus in a cardiovascular system the blood pressure input to an artery can be considered a train of periodic pulses and the blood pressure output is a periodic function of time. Both input and output can be recorded by means of the probes and the problem is to identify the artery transfer function as a linear time invariant model. In the engineering systems the input may be in the form of (1) small periodic changes in the set points of the fuel to the boiler, (2) the throttle of an airplane while the output may be the temperature changes in the outgoing steam or (3) the speed changes of the airplane. In each case the problem boils down to the transfer function identification from the given input and output data.

The proposed scheme is applicable when the input and the output is periodic as well as aperiodic. Both the input function $f_I(t)$ and the output function f.(t) are considered as the responses of two fictitious linear time invariant systems G_I ( ) and G ( ), to which a test input t(t) is applied (see figure 1).

For the aperiodic case: $\left(f_I(t),\ f_o(t)\ \text{aperiodic}\right)$

$t\ (t) = $ unit impulse

$T\ (s) = 1$

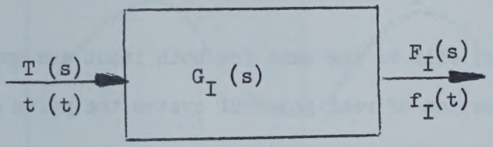
For the periodic case: $\left(f_I(t),\ f_o(t)\ \text{periodic}\right)$

$t\ (t) = $ a train of impulses T Seconds apart

$$T\ (s) = \frac{1}{1 - e^{-st}}$$

T is the time period of the functions $f_I(t)$ and $f_o(t)$. For the periodic case we will assume that the steady state is reached before the data is collected.

Input Function Synthesis



Output Function Synthesis

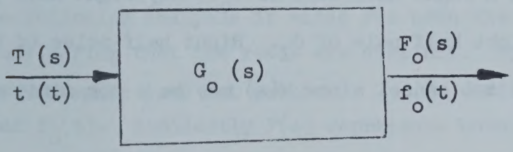


FIGURE 1.

The fictitious G's are assumed to have the general form

$$G\ (s) = K \frac{\left(s-z_1\right)\left(s-z_2\right)\ldots\qquad\left(s-z_m\right)}{\left(s-p_1\right)\left(s-p_2\right)\ldots\qquad\left(s-p_n\right)} \tag{1}$$

where $m < n$

where K is the gain

z's are zeros

p's are poles

Since these G's are fictitious, the poles and zeros need not occur in complex pairs and may be anywhere in the complex plane.

The transfer function of the physical system which is to be identified is

$$W(s) = \frac{F_O (s)}{F_I (s)} = \frac{G_O (s)}{G_I (s)}, \tag{2}$$

because for any given case T(s) will be the same for both input and output. Since W(s) is the transfer function of real physical system the poles must lie in the left half plane and complex poles and zeros must occur in conjugate pairs. This means that any non-conjugate pole or zero of G_O must cancel with a non-conjugate pole or zero of G_I. Similarly any right half zero of G_I must cancel with a right half zero of G_O. Any right half pole of G_O must cancel with a right half pole of G_I. Right half poles of G_I or right half zeros of G_O need not cancel since W(s) may be a non minimum phase transfer function.

The degree, m, of the numerators of the fictitious transfer functions G(s) must be less than the degree of , n, of the corresponding denominators. In case the functions f(t)'s are periodic, then if m = n-1, discontinuities occur in f(t)'s and all the derivatives at the start of each impulse of the test function.

If m = n-2 f(t)'s are continuous but discontinuities occur in the first and all higher derivatives at the impulse points and so on. These discontinuities if they are observable are a great help in determining the period and the impulse points of recorded data. However, in practice noise usually masks the discontinuities so that the impulse points cannot be determined even though the period is detectable.

In figure 2 the impulse points are obviously at 0, T, 2T, etc. However, due to the noise this might not be apparent in a real case so that the data may be analyzed between some instant τ and $\tau + T$. This is the reason for non-physically realizable G's, event though W(s) is physically realizable.

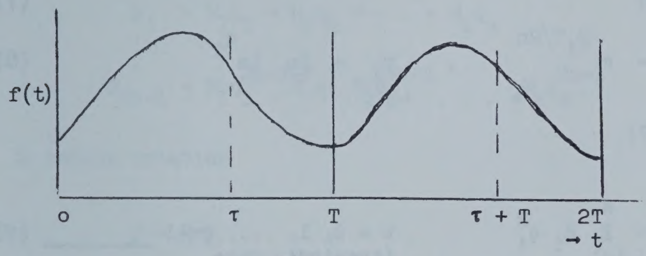


Figure 2

II OUTLINE OF ANALYSIS

The following analysis is valid for both the input $f_I(t)$ and the output $f_o(t)$ considering that the roots are distinct. The case of repeated roots can be handled, but space does not permit. In what follows f(t) represents both $f_I(t)$ and $f_o(t)$. Similarily F(s) represents both F_I and F_o.

For the aperiodic Case

$$F(s) = \sum_{i=1}^{n} \frac{R_i}{(s-p_i)} \qquad f(t) = \sum_{i=1}^{n} R_i e^{p_i t} \tag{3}$$

where

$$R_i = \Big[F(s)\,(s-p_i)\Big]_{s = p_i}, \qquad \text{the residue at the pole } p_i \tag{4}$$

For the Periodic Case

$$F(s) = \sum_{i=1}^{n} R_i \,|\, (s-p_i) \qquad f(\tau) = \sum_{i=1}^{n} R_i' e^{p_i \tau} \qquad o \leq \tau < T \tag{5}$$

where $$R_i' = \frac{R_i}{1 - e^{p_i T}} = \frac{R_i}{(1-q_i^{2n})}, \qquad \left(q_i = e^{p_i T/2n} \right) \tag{6}$$

Now suppose that f(t) is sampled at o, T/2n, ..., kT/2n, ..., (2n-1)T/2n.

For the periodic cast T is the period of the function f(t). For the aperiodic case the value of T should be large enough to include the significant variations in f(t).

Let

$$f\,\frac{(kT)}{2n} = f_k \tag{7}$$

$$q_i = e^{p_i T/2n} \quad , \quad p_i = \frac{2n}{T}\,\mathrm{Ln}\,q_i \tag{8}$$

From (3) and (7)

$$f_k = \sum_{i=1}^{n} R_i\, q_i^k \qquad k = 0, 1, \ldots, 2n\text{-}1 \;\underline{\hspace{2cm}} \quad \text{(Aperiodic Case} \tag{9}$$

or

$$f_k = \Sigma\, R_i^{'}\, q_i^k \qquad k = o, 1, \ldots, 2n - 1 \;\underline{\hspace{2cm}} \quad \text{(Periodic Case)} \tag{10}$$

Since the only difference between aperiodic and periodic situation is the interchange of R_i's with $R_i^{'}$'s, we will concentrate on the periodic case.

In order to make the analysis independant of the scaling of the measuring device it is convenient to normalize the measured samples f_k by dividing by the largest sample. This has the added advantage that if the process is programmed for a computer, the computer will have to deal with numbers less than or equal to unity.

Let $$y_k = \frac{f_k}{f_{max}} = \sum_{i=1}^{n} M_i\, q_i^k \tag{11}$$

where

$$M_i = \frac{R_i}{f_{max}} \tag{12}$$

Equation (11) represents an array of 2n equations with 2n unknowns, n M's and n q's

Thus

$$
\begin{aligned}
y_o &= M_1 + M_2 \ldots\ldots + M_n \\
y_1 &= M_1 q_1 + M_2 q_2 + \ldots + M_n q_n \\
&\vdots \\
y_{2n-1} &= M_1 q_1^{2n-1} + M_2 q_2^{2n-1} + \ldots + M_n q_n^{2n-1}
\end{aligned}
\tag{13}
$$

In matrix notation

$$\underset{\sim}{y} = \underset{\sim}{Q}\,\underset{\sim}{m} \tag{14}$$

where

$$\underset{\sim}{y} = \begin{bmatrix} y_o \\ y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \underset{\sim}{m} = \begin{bmatrix} M_1 \\ M_2 \\ \cdot \\ \cdot \\ \cdot \\ Mn \end{bmatrix} \quad \underset{\sim}{Q} = \begin{bmatrix} 1 & 1 \;\ldots & 1 \\ q_1 & q_2 & q_n \\ \cdot & & \\ \cdot & & \\ q_1^{2n-1} & q_2^{2n-1} & q_n^{2n-1} \end{bmatrix} \tag{15}$$

Let us define:

$$
\begin{aligned}
U_1 &= (-1) \; (q_1 + - - - q_n) \\
U_2 &= (-1)^2 \,(q_1 q_2 + q_1 q_3 - - - + q_2 q_3 - - q_n q_{n-1}) \\
&\vdots \\
U_n &= (-1)^n \;(q_1 q_2 - - - q_n)
\end{aligned}
\tag{16}
$$

The U's are the well know symmetric funcions.

With the help of (16) the M's can be eliminated from equations (14) yielding:

$$
\begin{aligned}
&U_n y_o + U_{n-1} y_1 + \cdot \;\; \cdot \;\; \cdot \;\; U_1 y_{n-1} + y_n = 0 \\
&\vdots \\
&U_n y_{n-1} + U_{n-1} y_n + \cdot \;\; \cdot \;\; \cdot \;\; U_1 y_{2n-2} + y_{2n-1} = 0 \\
&U_n + U_{n-1} q + \cdot \;\; \cdot \;\; \cdot \;\; U_1\, q^{n-1} + q^n = 0
\end{aligned}
\tag{17}
$$

This is an array of n + 1 equations and only n unknowns. Therefore:

$$D = \begin{bmatrix} y_o & y_1 & \text{----} & y_n \\ y_1 & y_2 & \text{----} & y_{n+1} \\ & & & \\ y_{n-1} & y_n & \text{----} & y_{2n-1} \\ 1 & q & & q^n \end{bmatrix} = 0 \tag{18}$$

exapanding this determinant in terms the last row yields a polynominal

$P(q)$ of degree n in q whose roots are $q_1, q_2, \ldots, q_n$

$$P(q) = \Delta_o + \Delta_1 q + \Delta_2 q^2 + \ldots + \Delta_n q^n = 0 \tag{19}$$

where the Δ's are the co-factors corresponding to the last row of the determinant i equation (18).

Dividing (19) by Δ_n we get

$$A_o + A_1 q + A_2 q^2 - - - A_{n-1} q^{n-1} + q^n = 0 \tag{20}$$

where $A_i = \dfrac{\Delta_i}{\Delta_n}$

Solution of the roots of (20) yield the various q's. The various poles p's may be determined from q's from (8). Notice that the poles have been determined independently of their residues (or zeros). The residues are obtained as follows:

Considering only the alternate equations from (13), we get

$$\underset{\sim}{y}_1 = \underset{\sim}{V}\,\underset{\sim}{m} \tag{21}$$

where

$$\underset{\sim}{y_1} = \begin{bmatrix} y_o \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_{2n-2} \end{bmatrix} \qquad \underset{\sim}{m} = \begin{bmatrix} M_1 \\ M_2 \\ \cdot \\ \cdot \\ \cdot \\ M_n \end{bmatrix} \qquad \underset{\sim}{V} = \begin{bmatrix} 1 & 1 & & 1 \\ q_1^2 & q_2^2 & & q_n^2 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ q_1^{2n-2} & q_2^{2n-2} & & q_n^{2n-2} \end{bmatrix} \tag{22}$$

The matrix V is the well known Vander Monde matrix whose inverse is easy to compute.

Thus $\underset{\sim}{m} = \underset{\sim}{V}^{-1} \underset{\sim}{y}_1$

And Residues are:

$$R_i = f_{max} M_i \quad (\text{ aperiodic Case }) \tag{24}$$

$$R_i = f_{max} M_i (1-q^{2n}) \quad (\text{ periodic Case }) \tag{25}$$

Any n equations from (13) can be used to evaluate the residues Ri. However, the use of every other equation utilizes evenly spread data and yields satisfactory answer. One can utilize various sets of n equations and thus obtain sets of residues which can be averaged, thereby reducing the effects of inaccuracies.

III. DETERMINATION OF ORDER n OF THE SYSTEM AND EFFECT OF NOISE (Measurement Inaccuracies)

In all the analysis so far n has been assumed to be known. However, in practice n is unknown. It turns out that if n is chosen larger than the actual order of the fictitious G then with perfect data all the Δ's will be identically zero, so the order of n could be lowered till non-zero Δ's are obtained. However, in practice the data is contaminated with noise. Thus, largerthan actual n will yield Δ's which are small but not zero. The question then arises, how small these Δ's should be before a decision is made to go to a smaller n?

Two schemes are proposed to answer this question:

1. Assume an n larger than the expected order of G. Then compute the Δ's. If all Δ's are < some number δ, go to a smaller n. The value δ will depend on the expected error in the measurements and the value n.

2. The second method assumes that more samples are taken than are needed

for the highest n anticipated. For example, assume that the highest n is 6. Then take 120 samples. The number 120 is evenly divisible by 2,4,6,8,10, and 12. These are the values of 2n when n goes from 1 to 6. Now assume first that n = 1. Choose samples 0 and 60 and compute Ao. Do the same for samples 6 and 66, 12 and 72, and so on up to 54 and 114. This will give ten values of Ao in all. If n = 1 is the correct value, all these Ao will be nearly equal (they would be exactly equal except for noise). If they do not vary more than say 10% from the average value, $\bar{A}o$, then use the average value in equation 18 which will be

$$q + \bar{A}o = 0 \quad \text{or } q = -\bar{A}o \tag{26}$$

If they do vary more than 10% from the average, go to n = 2. Now compute the A's from 10 groups of 4 samples say 0, 30, 60, and 90; 3, 33, 63, and 93 and so on up to 27, 57, 87, and 117. This will yield 10 Ao's and 10 A_I's. If no individual A varies from the average by more than 10% then $q_{1.}$ and q_2 will be the roots of the equation

$$q^2 + \bar{A}_1 q + \bar{A}_o = o \quad \text{(bar above A indicates average)} \tag{27}$$

Keep going in this way for third, forth, fifth and sixth degree polynomials, if necessary. If there is still too much variation when the sixth degree is reached, it would be necessary to start over with more samples However, if a fit cannot be obtained with as many as six poles, it would be questionable whether this method has much value.

RESPONSE OF SAMPLED-DATA-HOLD SYSTEMS BETWEEN SAMPLING INSTANTS

WILLIE L. McDANIEL, JR.
Mississippi State University, State College, Mississippi

ABSTRACT

The x-transform is a new transform which has been developed for application to sampled-data-systems which have each sampler followed by a zero-order hold device. The x-transform is defined as

$$\mathcal{X}[e(t)]=E(x)=E(s)\Big|_{(e^{-nTs}/s)=x^n}$$

where T is the sampling period in seconds, s is the Laplace variable and n=0, 1, 2---. The x-transform parallels the z-transform through basic definitions and theorems.

Two methods for determining the system behavior between sampling instants are given for both open- and closed-loop configurations. The methods are submultiple sampling and the modified x-transform.

Several manipulations in "x-algebra" must be mastered in order to become proficient with the x-transform method. Except for these few manipulations, the method is straightforward and applies to all sampled-data systems having each sampler followed by a zero-order hold.

The general equations for the submultiple sampling method are:

Open-loop

$$C(x_n)=E(x_n^n)G_R(x_n)$$

$$G_R(x_n)=G_R(x)\Big|_{\substack{x\to x_n\\ T\to T/n}}$$

$$E(x_n^n)=E(x)\Big|_{x\to x_n^n}$$

Closed-loop (Sampler-hold in the error path)

$$C(x_n)=G_R(x_n)\,\frac{U(x_n^n)}{1+GH_R(x_n^n)}$$

$$GH_R(x_n^n)=GH_R(x)\Big|_{x\to x_n^n}$$

$$GH_R(x)=\mathcal{X}\left[\frac{G(s)\;H(s)}{s}\right]$$

The general equations for the modified x-transform method are:

Open-loop

$$C(x,m)=E(x)G_R\,(x,m) \qquad [0\le m\le 1]$$

$$G_R(x,m)=x\sum_{n=0}^{\infty}g_R[(n+m)T](x^n-x^{n+1})$$

Closed-loop (Sampler-hold in the error path)

$$C(x,m)=G_R(x,m)\,\frac{U(x)}{1+GH_R(x)}$$

AN APPROXIMATE IDENTITY OPERATOR FOR CONTINUOUS SERVOMECHANISMS WITH TIME LAG

GLEN H. FOUNTAIN
Applied Physics Laboratory
Silver Spring, Maryland

CHARLES A. HALIJAK
University of Denver
Denver, Colorado

A method of control system design based upon admissibility of the control system into the servo class is presented. A criterion for admissibility of control systems with time lag is described. This criterion, based upon King's zero error coefficient theory, is expressed as a set of relations between coefficients of the control system transfer function.

Using Rault's work as a starting point, a method of realizing an improvement in the order of aidentity is presented. This development gives a systematic design for control systems.

Consider a transfer function of the form

$$\frac{\sum\limits_{n=0}^{N}\sum\limits_{m=0}^{N} a_{n,m}\, s^{n}\, e^{-mTs}}{\sum\limits_{n=0}^{N}\sum\limits_{m=0}^{N} b_{n,m}\, s^{n}\, e^{-mTs}} = \frac{\sigma' A \zeta}{\sigma' B \zeta} \equiv M(s,z)$$

where

$\sigma = (s^0, s^1, s^2, \ldots, s^N)'$

$\zeta = (z^0, z^1, z^2, \ldots, z^N)'$

$Z \equiv e^{-Ts}$

A and B are square matrices of order n+1

prime denotes the transpose operation.

Futhermore, let a matrix be defined by

$$\underset{(N+1)\times(N+1)}{Q_k} = \begin{pmatrix} o' & v' \\ i^{k-j} & uv' \end{pmatrix}$$

where

i ranges from 1 to N

j ranges from 1 to k-1

$v = (1, 0, 0, \ldots, 0)'$, an $(N+2-k)$ - vector

$u = (1, 1, 1, \ldots, 1)'$, an $(N+2-k)$ - vector

The aidentity (approximate identity) criterion for the rational bilinear form can be reduced to an algorithm of the form

$$\text{diagonal } (AQ_k) = \text{diagonal } (BQ_k)$$
$$k = 1, 2, 3, \ldots, n$$

An additional technique is presented for generating stable aidentities.

ON THE ABSOLUTE STABILITY OF SYSTEMS CONTAINING MULTIPLE NONLINEARITIES

M.D.SRINATH, M.A.L.THATHACHAR AND H.K.RAMAPRIYAN
Department of Electrical Engineering, Indian Institute of Science, Bangalore - 12, India.

ABSTRACT

A Lyapunov function of the quadratic plus multi-integral type has been used by other authors for obtaining frequency-domain criteria for the absolute stability (asymptotic stability in the large) of autonomous systems containing multiple nonlinearities of monotone and odd-monotone type by utilizing a matrix version of the Kalman lemma. These results are restricted to the direct control case and make use of only the real zeros of the linear part transfer functions.

In the present paper a method of taking into account complex zeros by considering other subclasses of monotone increasing nonlinearities is given and both direct and indirect control cases are considered . In essence the technique used in the paper involves changing the zeros of the linear part transfer function matrix so that a frequency domain inequality is satisfied. Restrictions on the location of zeros to be replaced as well as on the introduction of new zeros can be lessened as the nonlinear function approaches linearity and thus stability can be assured for a larger class of systems.

A GENERALIZED ZUBOV FORMULATION FOR LIMIT CYCLES IN THIRD ORDER NONLINEAR SYSTEMS

M. J. WOZNY
Department of Electrical Engineering
Purdue University
Lafayette, Indiana

D. G. SCHULTZ
Department of Electrical Engineering
University of Arizona
Tucson, Arizona

ABSTRACT

Limit cycle behavior in third order systems is studied by first constructing a Liapunov function, V, which represents a toroid, and then by examining $\dot{V}$ to show that all trajectories crossing the toroidal surface pass into its interior and there tend to the limit cycle. Such an approach requires that the Liapunov function representing the toroid be general enough to make it applicable to the various motions possible in third order nonlinear systems.

A Liapunov formulation is presented in which the toroid is generated via two unknown surfaces whose intersection represents the system limit cycle. Using this interpretation it is shown that Liapunov functions considered previously are too restrictive. This restriction is removed, formally at least, and an example is given to illustrate the wider class of systems permitted in the new formulation.

1. INTRODUCTION

In three dimensional limit cycle problems, the Poincare-Bendixson Theorem requires that a toroid be constructed which contains no singular points such that trajectories only enter the toroid and there circle within it.(1) This paper is concerned with the formulation of this theorem via the Second Method of Liapunov.

Consider the requirements of the Liapunov formulation. Given a Liapunov function V which represents a toroid, the question of whether the trajectories enter the toroid is determined by examining $\dot{V}$. To insure that the trajectories circle within the toroid and tend to the limit cycle, the Liapunov function is required to represent a nested set of toroids which close down on the limit cycle as V approaches zero. Finally, $\dot{V}$ must be zero on the limit cycle to insure that it is a trajectory of the system. These conditions indicate that the construction of the toroid must somehow characterize the unknown limit cycle exactly.

---

* This work was supported by the National Science Foundation under Grant GP-2237 and by the National Aeronautics and Space Administration under Grant NGR 15-005-021.

Indeed, Zubov(4) used such a formulation when he demonstrated that the region of asymptotic stability of the limit cycle

$$x_1^2 + x_2^2 = 1 \qquad (1)$$

$$x_3 = 0$$

of the system

$$\begin{aligned} \dot{x}_1 &= x_1 + x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 &= -x_1 + x_2 - x_2(x_1^2 + x_2^2) \qquad (2) \\ \dot{x}_3 &= -x_3 \end{aligned}$$

is the entire E^3-space, excluding the x_3-axis. In particular, he defined the Liapunov function

$$V = \frac{(1 - x_1^2 - x_2^2)^2}{x_1^2 + x_2^2} + x_3^2 \qquad (3)$$

which for every V = constant represents a toroid enclosing the limit cycle. All trajectories, except those originating on the x_3-axis, enter this toroid since

$$\dot{V} = -\frac{2(1 + x_1^2 + x_2^2)(1 - x_1^2 - x_2^2)^2}{x_1^2 + x_2^2} - 2x_3^2 \qquad (4)$$

is negative definite with respect to the limit cycle. The system behavior is indicated in Fig. 1.

Szegö and Geiss(3) considered the problem of finding a stable n-1 dimensional surface, S, in E^n via the Second Method of Liapunov. Their approach can be summarized as follows. If there exists a Liapunov function $V(\underline{x})$ with

$$V(\underline{x}) = 0 \text{ for } \underline{x} \in S \qquad (5)$$

such that

$$\dot{V} = -\Theta(V) \qquad (6)$$

where

$$V\Theta(V) \geqslant 0 \qquad (7)$$

and

$$\Theta(V(\underline{x})) = 0 \text{ for } x \in S \tag{8}$$

$$\Theta(V(\underline{x})) \neq 0 \text{ for } x \notin S$$

then the surface S is asymptotically stable.

When applied to limit cycle problems in E^3, this approach requires that there exist an asymptotically stable surface containing the limit cycle. If such is the case, then the problem can be reduced to a two dimensional one by examining the system behavior constrained to this surface.

In the next section, a Liapunov formulation is given which extends these ideas, and hence, is applicable to a larger class of systems.

2. FORMULATION

The Poincaré-Bendixson Theorem is reformulated in terms of the Second Method of Liapunov. By defining the toroid in terms of two surfaces, a general approach is formulated for studying limit cycles in E^3. The generality of this approach is discussed by comparing it to existing methods.

Let

$$\dot{\underline{x}} = \underline{f}(\underline{x}), \ \underline{f}(\underline{0}) = \underline{0} \tag{9}$$

be a third order system having one limit cycle. Since the concept of Liapunov functions is applied to equilibrium states other than singular points, the definition of positive definiteness must also be considered in a broader sense.

Definition 1

A function $V(\underline{x})$ is said to be positive definite with respect to a closed curve, C, if in the neighborhood of C,

$$V(\underline{x}) = 0 \text{ for } x \in C \tag{10}$$

$$V(\underline{x}) > 0 \text{ for } x \notin C.$$

The Liapunov formulation of the limit cycle problem can be stated as follows.

Theorem 1

If there exists a Liapunov function $V(\underline{x})$ for system (9) and a closed curve, C, satisfying the conditions:

(1) $V(\underline{x})$ = constant represents a toroid about C,
(2) $V(\underline{x})$ is positive definite with respect to C,
(3) $\dot{V}(\underline{x})$ is negative definite with respect to C,

then C is the limit cycle of the system.

The proof of this theorem follows directly from the Poincare-Bendixson Theory given previously and from LaSalle's interpretation of Liapunov stability.(2)

More structure is now added to Theorem 1 by letting the Liapunov function be characterized by a particular representation for the toroid. Let a closed curve, C, be defined by the intersection of the two surfaces

$$S(\underline{x}) = 0 \qquad (11)$$

$$W(\underline{x}) = 1.$$

Next define a Liapunov function

$$V(W(\underline{x}),S(\underline{x})) \qquad (12)$$

in terms of these surfaces such that it is positive definite with respect to C, and such that

$$V = \text{constant} \qquad (13)$$

represents a toroid about C. Since V is positive definite with respect to C, then for successively smaller values of the constant, Eq. (13) represents a nested set of toroids which close down to C.

Differentiating Eq. (12) with respect to the independent variable t, form

$$\dot{V} = \frac{\partial V}{\partial W}\dot{W} + \frac{\partial V}{\partial S}\dot{S}. \qquad (14)$$

To insure that trajectories cross into the toroid, choose

$$\dot{W} = F(\underline{x}) \qquad (15)$$

$$\dot{S} = G(\underline{x})$$

such that

$$\dot{V} = \frac{\partial V}{\partial W}F(\underline{x}) + \frac{\partial V}{\partial S}G(\underline{x}) \qquad (16)$$

is negative definite with respect to C.

This approach is summarized in the next theorem.

Theorem 2

If there exist functions $W(\underline{x})$, $S(\underline{x})$ satisfying (15), then the closed curve C, given by (11), is the limit cycle of system (9).

Consider next the Zubov example interpreted in terms of Theorem 2.

In the Zubov example, the S = 0 surface of Eq. (11) is the x_1,x_2-plane, $x_3 = 0$, and the W = 1 surface is the right cylinder, $x_1^2 + x_2^2 = 1$. Clearly, the intersection of these surfaces generate the desired limit cycle.

Furthermore,

$$\dot{W} = 2(x_1^2 + x_2^2)(1 - x_1^2 - x_2^2) \qquad (17)$$
$$\dot{S} = -x_3.$$

Eqs. (17) indicate that any trajectory originating above the x_1,x_2-plane, i.e., for $x_3 > 0$, will tend to the x_1,x_2-plane from the $x_3 > 0$ direction regardless of what other motion it it may exhibit in the x_1 and x_2 directions.

This decoupling effect between the W and S surfaces eliminates from consideration, in both the Zubov and Szegö formulations, the class of systems whose trajectories spiral about the limit cycle as they tend to it. Conversely, by allowing coupling to exist, the Liapunov formulation of Theorem 2 is applicable to a wider class of systems.

An example of a system whose trajectories spiral about the limit cycle is given in the next section.

3. EXAMPLE

Consider the following third order system.

$$\begin{aligned}
\dot{x}_1 &= x_1 + x_2 - x_1(x_1^2 + x_2^2) + x_3(x_1^2 + x_2^2)^2 \\
\dot{x}_2 &= -x_1 + x_2 - x_2(x_1^2 + x_2^2) + x_3(x_1^2 + x_2^2)^2 \qquad (18) \\
\dot{x}_3 &= x_1 + x_2 - x_3 - (x_1 + x_2)(x_1^2 + x_2^2)
\end{aligned}$$

This system was chosen because its behavior is similar to the Zubov example. Following a generalized version of Zubov's approach, let the Liapunov function be chosen as

$$V = \frac{(1 - W)^2}{W} + S^2 . \qquad (19)$$

Clearly

$$V(W,S) = 0 \qquad (20)$$

for

$$W = 1 \qquad (21)$$
$$S = 0$$

i.e., on the closed curve, C. Next force

$$\dot{V} = -\frac{(1 + W)(1 - W)}{W^2}\dot{W} + 2S\dot{S} \qquad (22)$$

to be negative definite with respect to the closed curve C by making suitable choices for $\dot{W}$ and $\dot{S}$.

Allowing for coupling between the surfaces, let

$$\begin{aligned} \dot{W} &= 2W(1 - W) + p(\underline{x}) \\ \dot{S} &= -\theta(S) + q(\underline{x}) \end{aligned} \qquad (23)$$

where the coupling terms satisfy

$$\frac{(1 + W)(1 - W)}{W^2} p(\underline{x}) + 2Sq(\underline{x}) = 0 \qquad (24)$$

and

$$S\theta(S) \geqslant 0 \qquad (25)$$

in the sense of the Szegö formulation. Then

$$\dot{V} = - \frac{2(1 + W)(1 - W)^2}{W} - 2S\theta(S) \qquad (26)$$

is negative definite with respect to C.

It can be shown that if the following choices are made

$$\begin{aligned} \theta(S) &= S \\ p(\underline{x}) &= 2W^2 S(x_1 + x_2) \\ q(\underline{x}) &= (1 + W)(1 - W)(x_1 + x_2) \end{aligned} \qquad (27)$$

then a solution exists for Eqs. (23) and is given by

$$\begin{aligned} W(\underline{x}) &= x_1^2 + x_2^2 \\ S(\underline{x}) &= x_3 \ . \end{aligned} \qquad (28)$$

Thus the limit cycle is defined by

$$\begin{aligned} x_1^2 + x_2^2 &= 1 \\ x_3 &= 0. \end{aligned} \qquad (29)$$

The behavior of this system is illustrated in Fig. 2 where it is seen that the trajectories spiral about the limit cycle as they tend to it.

4. CONCLUSIONS

In this paper a Liapunov formulation was proposed for determining the behavior of limit cycles in third order nonlinear systems. It was shown that previous formulations of this problem are valid only for a restrictive class of systems. This restriction was removed in the proposed formulation by allowing coupling terms to exist in the defining equations of the surfaces generating the toroid. An example illustrated this general behavior. No attempt was made to present a general method for generating Liapunov functions required by this formulation. This is still an area for further research.

REFERENCES

(1) Friedrichs, K.O., "On Nonlinear Vibrations of Third Order," Studies in Nonlinear Vibration Theory, New York University, 1946, 65-103.

(2) LaSalle, J.P., "An Invariance Principle in the Theory of Stability," Technical Report 66-1, Center for Dynamical Systems, Brown University, April, 1966.

(3) Szegö, G.P. and G.R. Geiss, "A Remark on 'A New Partial Differential Equation for the Stability Analysis of Time Invariant Control Systems'," J. SIAM Control, Series A, Vol. 1, No. 3 (1963).

(4) Zubov, V.I., Mathematical Methods for the Study of Automatic Control Systems. New York: Pergamon Press, 1963, 203.

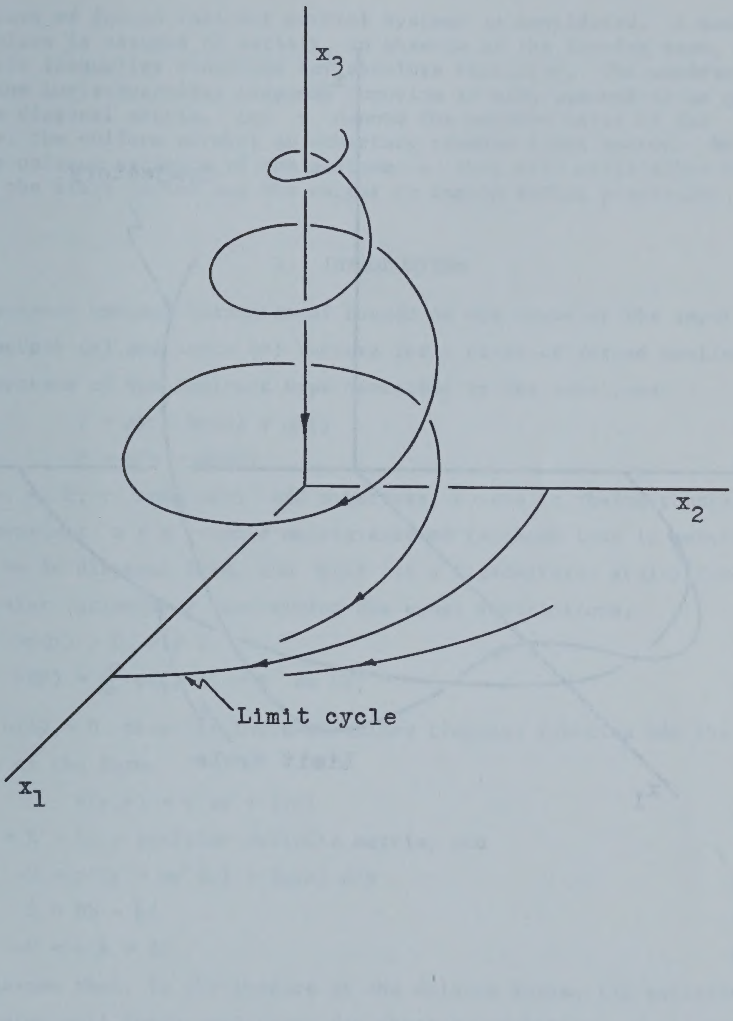

x_3
x_2
x_1
Limit cycle

Figure 1 Typical Behavior of System (2)

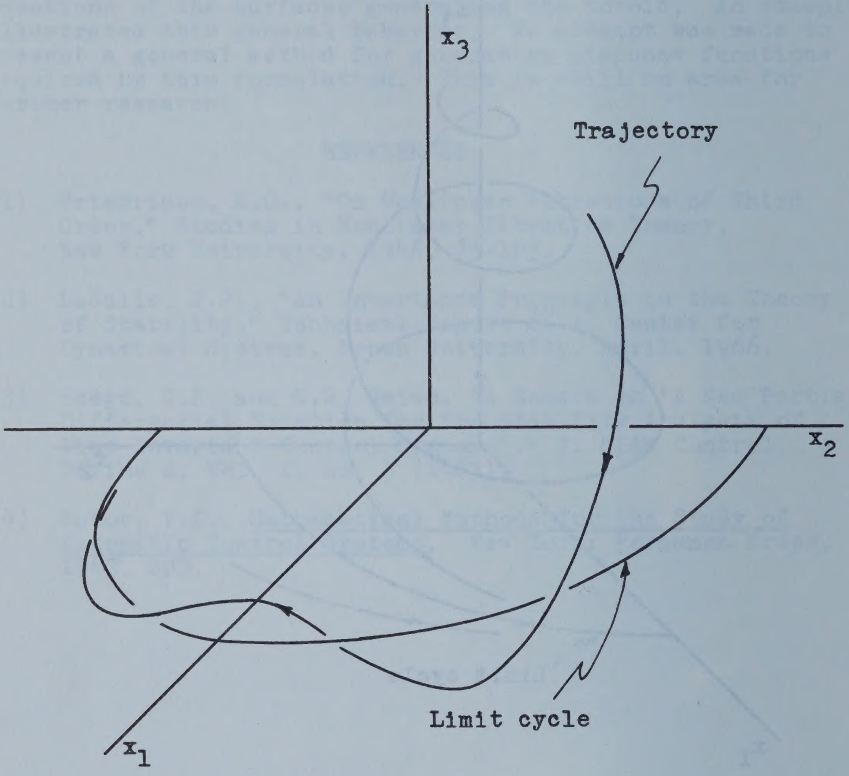

x3
Trajectory
x2
Limit cycle
x1

Figure 2 Typical Behavior of System (18)

INPUT AND OUTPUT BOUNDS FOR FORCED NONLINEAR CONTROL SYSTEMS*

RUI J. P. DE FIGUEIREDO and MICHEL M. DE VRIES**
Rice University, Houston, Texas, 77001

ABSTRACT

A class of forced indirect control systems is considered. A member of this class is assumed to satisfy, in absence of the forcing term, a Lurie matrix inequality condition for absolute stability. The quadratic form in the Lurie-Postnikov Liapunov function is also assumed to be generated by a diagonal matrix. Let ν denote the maximum value of the norm (i.e. the uniform norm) of an arbitrary bounded input vector. We obtain an optimum estimate of the maximum ν that will still allow the norms of the state vector and the output to remain within prescribed bounds.

1. INTRODUCTION

We present optimum estimates of bounds on the norms of the input $(u(t))$, output (σ) and state (y) vectors for a class of forced nonlinear control systems of the indirect type described by the equations:

$$\dot{y} = Ay - b\varphi(\sigma) + u(t) \tag{1a}$$

$$\dot{\sigma} = c'y - \rho\varphi(\sigma) \tag{1b}$$

Here, y, b, c and u(t) are n-vectors, b and c being constant , A is a constant $n \times n$ stable matrix assumed (without loss in generality) to be in diagonal form, and $\varphi(\sigma)$ is a Lipschitzian scalar function of the scalar variable σ satisfying the usual restrictions:

$$\sigma\varphi(\sigma) > 0, \ \sigma \neq 0 \ , \tag{2a}$$

$$\Phi(\sigma) = \int_0^{\sigma} \varphi(\zeta)\, d\zeta \rightarrow \infty \ \text{ as } |\sigma| \rightarrow \infty \tag{2b}$$

If $u(t) = 0$ then the Lurie-Postnikov Liapunov function for the system is of the form

$$V(y,\sigma) = y'By + \Phi(\sigma) \tag{3}$$

where $B = B'$ is a positive definite matrix, and

$$-\dot{V} = y'Cy + \rho\varphi^2(\sigma) + 2\varphi(\sigma)\, d'y \tag{4}$$

$$d = Bb - \tfrac{1}{2}c \tag{5}$$

$$-C = A'B + BA \tag{6}$$

We will assume that, in the absence of the driving force, (1) satisfies the following well-known conditions for absolute stability: C is positive definite matrix and

$$\rho > d'C^{-1}d \tag{7}$$

---

*This research was supported by the National Science Foundation under Grant GU-1153.

**Teagle Foundation Fellow at Rice University.

Finally, we will restrict our attention to only those cases where C and hence B is in a diagonal form. The diagonal elements in the matrix C (and hence B) will be adjustable.

Under these conditions we obtain the aforementioned bounds in the context of the following problem.

How large can the excursion ν in the norm of an arbitrary bounded input $u(t)$ to the system (1) be and still allow the excursions in the norms of the state vector y and input σ to remain within prescribed bounds?

To the above, there corresponds the following dual problem: Suppose it is known that the excursions in the norms of the inputs to the system will not exceed a certain range. Find the best (lowest) possible bounds on the corresponding excursions in $\|y\|$ and $|\sigma|$.

The practical significance of both these problems in the design and operation of nonlinear control systems is quite evident and so it needs no elaboration.

We will only consider the first problem since an analytical answer to it also constitutes a solution to the dual problem. The approach taken by us in obtaining the desired result will now be described.

2. DESCRIPTION OF THE METHOD

For a given t_o and an arbitrary bounded $u(t)$ define

$$\nu = \max_{t \geq t_o} \|u(t)\| \tag{8}$$

Let there be given two positive constants ϵ and σ_M which define the region Ω in the $(y\text{-}\sigma)$ space

$$\Omega = \{(y,\sigma) : |y_i| \leq \epsilon \text{ , } i = 1, \ldots, n, |\sigma| \leq \sigma_M \} \tag{9}$$

inside which the solution $(y(t), \sigma(t))$ of (1) must lie for all $t \geq t_o$. Our problem then consists in estimating the largest ν for which this is guaranteed.

The starting point will be a general formula for ν given by Malkin [1] for a system of the form*

$$\dot{x} = f(x) + v(t) \tag{10}$$

where f is a Lipschitzian m-vector function of the m-vector x and the m-vector $v(t)$ represents the forcing term. Malkin's result may be expressed as follows:

---

*Although Malkin's result is usually stated for the more general system $\dot{x} = f(x,t) + v(x,t)$, for the purposes of the present paper it will be sufficient to restrict it to systems of the form (10).

Suppose the null solution of the unforced system $\dot{x} = f(x)$ is globally asymptotically stable with a Liapunov function $V(x)$. Given an arbitrary positive constant ϵ, let K = minimum of $V(x)$ on the surface $\|x\| = \epsilon$, K' = maximum of $-\dot{V}(x) = -f(x) \cdot \text{grad } V$ on the surface $V(x) = K$, and $\|x\| = \epsilon_1$ the largest spherical surface in the region $\{x : -\dot{V}(x) \leq K'\}$. For a given t_o and x_o let $x(t, x_o, t_o)$ denote a solution of (10) satisfying $x(t_o, x_o, t_o) = x_o$. Then for an arbitrary t_o and for every x_o such that $\|x_o\| < \epsilon_1$, we have $\|x(t, x_o, t_o)\| < \epsilon$ for all $t \geq t_o$ provided

$$\max_{t \geq t_o} \|v(t)\| \leq \nu = \frac{\alpha K'}{mM} \tag{11}$$

where $\alpha = 1 - \alpha_o$, where α_o is an arbitrarily small positive constant and $M = \max \left|\frac{\partial V}{\partial x_i}\right|$, $i = 1, \ldots, m$, in the region $\|x\| \leq \epsilon$.

In applying this result to our case, we identify (10) with (1) by letting $m = n + 1$, $x = (y,\sigma)$ and $v(t) = (u(t),0)$, and obtain the parameters in (11) by using the expressions for $V(y,\sigma)$ and $V(y,\sigma)$ in (3) and (4). However, it is important to note that the value of ν in (11) depends on the choice of the V function. Hence an analytical expression for ν must be obtained in terms of the adjustable constants (B matrix) in the V function in order that the result may be optimized for maximum ν. But leaving the constants in V arbitrary and the weak restrictions on the functional dependence of φ on σ assumed do not permit the derivation of an exact analytical expression for K' in (11).

For this reason we obtain an approximate expression for ν constructed in the way indicated below. We will show in section 4 that the V function that optimizes this expression is the same as the one described as the optimum V function for the unforced case by LaSalle and Lefschetz [2, 3].

Let

$$A = \text{diag}(\lambda_1, \ldots, \lambda_n) \tag{12}$$

$$\lambda_i = -\mu_i + j\omega_i, \quad i = 1, \ldots, n, \ \mu_i \text{ and } \omega_i \text{ real and } \mu_i > 0 \tag{13}$$

$$B = \text{diag}(B_{11}, \ldots, B_{nn}), \quad B_{ii} > 0, \ i = 1, \ldots, n \tag{14}$$

$$C = \text{diag}(C_{11}, \ldots, C_{nn}), \quad C_{ii} > 0, \ i = 1, \ldots, n \tag{15}$$

and by (6)

$$B_{ii} = \frac{C_{ii}}{2\mu_i}, \qquad i = 1, \ldots, n. \tag{16}$$

Let

$$B_{kk} = \min (B_{11}, \ldots, B_{nn}) \quad . \tag{17}$$

We proceed along the following steps in obtaining ν :

(a) We choose the largest $V(y,\sigma)$ = constant surface S in Ω by letting

$$V(y,\sigma) = y'By + \Phi(\sigma) = \hat{\Phi}(\sigma_M) = K, \ \hat{\Phi}(\sigma_M) = \min\{\Phi(\sigma_M), \Phi(-\sigma_M)\}, \tag{18}$$

and adjusting later on the constant $B_{kk} = (2\mu_i)^{-1} C_{kk}$ to satisfy

$$C_{kk} = 2\mu_k K/\varepsilon^2 \quad . \tag{19}$$

Then S intersects the σ and y_k axes respectively at σ_M (or $-\sigma_M$) and another value of σ inside Ω and at $\pm \varepsilon$. It intersects the other axes inside Ω.

For a given constant $\dot{V}$ surface S' in the (y,φ) space

$$-\dot{V}(y,\varphi(\sigma)) = y'Cy + 2\varphi d'y + \rho\varphi^2 = K' , \tag{20}$$

we obtain the maximum and minimum values that φ and y_i, $i = 1, \ldots, n$, attain on this surface by studying successively its projections on (y_{n-1}, y_n), (φ, y_n), (φ, y_{n-1}), $\ldots$, (φ, y_1) planes. With these values we construct the smallest rectangular hyperparallelipiped π that contains (and is tangent to) S'. Let π' be the rectangular hyperparallelipiped in the (y,σ) space whose image in the (y,φ) space is π.

(b) By adjusting the constant K' we touch the corners of π' to the surface S described by (18). Such a π' is clearly the largest one in the region bounded by S. Hence this value of K' gives our best estimate of ν in (11) for a given choice of constants C_{ii}, $i = 1, \ldots, n$.

(c) Finally, we maximize the expression for ν thus obtained with respect to the constants C_{ii} that we are free to adjust. Surprisingly enough the C_{ii} then obtained correspond to those for optimum V for the stability of the unforced case as mentioned earlier.

3. BOUNDS ON CONSTANT $\dot{V}$ SURFACES

Let us then calculate the maxima and minima of φ and y_i, $i = 1, \ldots, n$, on the surfaces S' as indicated in step (a).

For this purpose, let us define the diagonal matrices

$$C_i = \mathrm{diag}(C_{ii}, C_{i+1,i+1}, \ldots, C_{nn}), \quad i = 2, \ldots, n, \tag{21}$$

and the column vectors

$$\hat{d}_i = \mathrm{col}(d_i, \ldots, d_n), \quad i = 2, \ldots, n, \tag{22}$$

$$\hat{y}_i = \mathrm{col}(y_i, \ldots, y_n), \quad i = 2, \ldots, n. \tag{23}$$

Then for a given value of y_1, say y_{10}, in the range of y_1 covered by S', the intercept of the hyperplane $y_1 = y_{10}$ and S' will be, according to (20) through (23), the surface described by

$$C_{11}\, y_{10} + 2\varphi d_1\, y_{10} + (\hat{y}_2{}'C_2\hat{y}_2 + 2\varphi\hat{d}_2{}'\hat{y}_2 + \rho\varphi^2 - K') = 0 \, . \tag{24}$$

Clearly, if y_{10}^{2} corresponds to the maximum or minimum of y_1 on S', the two roots of the above quadratic equation in y_{10} are equal. To produce this condition we set the discriminant with respect to y_{10} in (24) equal to zero. This projects the surface S' into the manifold $(\varphi, y_2, \ldots, y_n)$, the lower order surface being defined by

$$-(1/C_{11})\varphi^2 d_1^2 + \hat{y}_2{}'Cy_2 + 2\varphi\hat{d}_2{}'\hat{y}_2 + \rho\varphi^2 - K' = 0 \, , \tag{25}$$

which is the same as

$$C_{22}y_2^2 + 2\varphi d_2 y_2 + (\hat{y}_3{}'C_3\hat{y}_3 + 2\varphi\hat{d}_3{}'\hat{y}_3 + \rho\varphi^2 - K' - (1/C_{11})\varphi^2 d_1^2) = 0. \tag{26}$$

Eliminating y_2 from (26) by a procedure similar to that used with (24) we obtain

$$C_{33}y_3^2 + 2\varphi d_3 y_3 + (\hat{y}_4{}'C_4\hat{y}_4 + 2\varphi\hat{d}_4{}'\hat{y}_4 + \rho\varphi^2 - K' - \varphi^2(\frac{d_1^2}{C_{11}} + \frac{d_2^2}{C_{22}})) = 0 \, . \tag{27}$$

Continuing in this way we arrive at

$$C_{nn}\, y_n^2 + 2\varphi d_n y_n + (-K' + \varphi^2(\rho - \sum_{i=1}^{n-1} d_i^2/C_{ii})) = 0 \, , \tag{28}$$

and finally (eliminating y_n by the discriminant procedure)

$$\varphi_m^2 = \max_{\varphi\in S'} \varphi^2 = \frac{K'}{\rho - \sum_{i=1}^{n} d_i^2/Cii} = \frac{K'}{\rho - d'C^{-1}d} \, . \tag{29}$$

If φ rather than y_n is eliminated from (28), we get

$$\eta_n^2 = \max_{y_n\in S'} y_n^2 = \frac{K'}{C_{nn}} \frac{(\rho - \sum_{i=1}^{n-1} d_i^2/C_{ii})}{(\rho - \sum_{i=1}^{n} d_i^2/C_{ii})} = \frac{K'}{C_{nn}} (1 + \frac{d_n^2/C_{nn}}{\rho - d'C^{-1}d}) \, . \tag{30}$$

In an analogous manner one shows that a similar formula applies to all the variables y_j, $j = 1, \ldots, n$:

$$\eta_j^2 = \max y_j^2 = \frac{K'}{C_{jj}}(1 + \frac{d_j^2/C_{jj}}{\rho - d'C^{-1}d}), \quad j = 1, \ldots, n. \tag{31}$$

It is clear that for each of the variables $\varphi, y_1, \ldots, y_n$, the maximum and minimum on S' have the same magnitude and opposite signs. The set of maxima and minima in these variables define the hyperparallelipiped

$$\pi = \{(\varphi, y_1, \ldots, y_n) : |\varphi| \leq \varphi_m, |y_j| \leq \eta_j, j = 1, \ldots, n\} . \quad (32)$$

Let us define the non-decreasing function

$$\hat{\varphi}(\sigma) = \begin{cases} \max\limits_{0 \leq \xi \leq \sigma} \varphi(\xi), & \sigma \geq 0 \\ \min\limits_{\sigma \leq \xi \leq 0} \varphi(\xi), & \sigma \leq 0 \end{cases} \quad (33)$$

Then every rectangular hyperparallelipiped in the (y,σ) space

$$\pi' = \{(\sigma, y_1, \ldots, y_n) : \sigma_{m1} \leq \sigma \leq \sigma_{m2}, |y_j| \leq \eta_j, j = 1, \ldots, n \text{ where } -\sigma_{m1}, \sigma_{m2} \text{ and } \eta_j \text{ are arbitrary positive constants}\} \quad (34)$$

maps into a unique hyperparallelipiped π in the (y,φ) space:

$$\pi = \{(\varphi, y_1, \ldots, y_n) : |\varphi|^2 \leq \max(\varphi^2(\sigma_{m1}), \varphi^2(\sigma_{m2})), |y_j| \leq \eta_j, j = 1, \ldots, n\} . \quad (35)$$

Clearly, the converse need not be true.

4. DETERMINATION OF THE CONSTANT K'

Let us now use these results to find the largest π' contained by S as outlined in part (b) of our procedure.

We define

$$\text{col } (\eta_1, \ldots, \eta_n) \quad (36)$$

where η_j, $j = 1, \ldots, n$, are as defined in (31).

Also, let $\Phi_+(\sigma)$ and $\Phi_-(\sigma)$ denote the restrictions of $z = \Phi(\sigma)$ to the half lines $[0, \infty)$ and $(-\infty, 0]$. Then since $\Phi(\sigma)$ is strictly monotonically increasing with $\pm\sigma$, $\Phi_+(\sigma)$ and $\Phi_-(\sigma)$ have unique inverses, say $\sigma_+ = \Phi_+^{-1}(z)$ and $\sigma_- = \Phi_-^{-1}(z)$.

It is clear that

$$\hat{\sigma}_{m1} \equiv \Phi_-^{-1}(K - \eta' B \eta) \quad (37a)$$

$$\hat{\sigma}_{m2} \equiv \Phi_+^{-1}(K - \eta' B \eta) \quad (37b)$$

exist for K' sufficiently small (see (18), (31) and (36)). For a given matrix B (or equivalently C), we obtain the desired value of K' corresponding to the largest π' bounded by S, by requiring that $\max\{\hat{\varphi}^2(\sigma_{m1}), \hat{\varphi}^2(\sigma_{m2})\} = \varphi_m^{\,2}$, where $\varphi_m^{\,2}$ is as in (29).

In fact, it follows from (31) and (16) that

$$\eta' B \eta = (1/2)K' \sum_{i=1}^{n} \frac{1}{\mu_i}\left(1 + \frac{d_i^{\,2}/C_{ii}}{\rho - d' C^{-1} d}\right) \quad (38)$$

So (in view of (29)) K' and the variable $z = \Phi(\sigma_{m1}) = \Phi(\sigma_{m2})$ are obtained by the simultaneous solution of the equations

$$K' = \frac{2(K - z)}{\sum_{i=1}^{n} \frac{1}{\mu_i}\left(1 + \frac{d_i^2/C_{ii}}{\rho - d'C^{-1}d}\right)} \quad , \quad z \geq 0 \tag{39}$$

$$\psi(z) = \max\{\hat{\varphi}^2(\Phi_-^{-1}(z)),\ \hat{\varphi}^2(\Phi_+^{-1}(z))\} = \frac{K'}{\rho - d'C^{-1}d} = \frac{2(K - z)}{\sum_{i=1}^{n} \frac{1}{\mu_i}\left(\rho - \sum_{\substack{k=1 \\ k \neq i}}^{n} d_k^2/C_{kk}\right)} \tag{40}$$

There reamins to be shown how to determine the matrix C (and hence B) so that K' may indeed be evaluated as above.

5. OPTIMUM CHOICE OF THE LIAPUNOV FUNCTION

With the substitution

$$w_i = d_i^2/C_{ii} \ , \quad i \ \ 1, \ldots, n \tag{41}$$

(40 may be written in the form

$$K - z = f(w_1, \ldots, w_n)\ \psi(z) \ . \tag{42}$$

Since $\psi(z)$ is non-decreasing, it is clear that the solution z of (42) (which may be visualized as the intercept of the graphs (K-z,z) and ($\psi(z)$,z)) has the highest value when the (parametric) function $f(w_1, \ldots, w_n)$ has the lowest value. Hence the values of $w_1, \ldots, w_n$ that minimize f are those that maximize (optimize) the range of z and hence of the output variable. It is readily seen from (40) that this requires w_i, $i = 1, \ldots, n$, and hence $d'C^{-1}d$ to have the lowest possible value. This is the same condition as that which gives the best diagonal C for the determination of the absolute stability of the unforced system from an entirely different set of considerations.

Finally, rewriting the first and second members of (40) in the form

$$K' = (\rho - d'C^{-1}d)\ \psi\ (z) \tag{43}$$

we note that the lowest values of w_i maximize both the factors in this expression (the reason for $\psi(z)$ being that it is non-decreasing with z).

Details on how the minimization of the w_i, $i = 1, \ldots, n$, is implemented according to the eigenvalues λ_i, $i = 1, \ldots, n$, of the A matrix are given in reference [3] and hence need not be repeated here.

6. CONCLUSION

Summarizing our procedure, in order to obtain an optimum (maximum value of the) bound ν on the norm of the input $u(t)$ which will allow $|\sigma(t)| \leq \sigma_M$ and $|y_i| \leq \epsilon$, $i \ \ 1, \ldots, n$, where σ_M and ϵ are given bounds:

(1) Obtain the components of the matrix C which minimize $d'C^{-1}d$ in the way indicated in reference [3]. Let them be $C_{ii} = \alpha\hat{C}_{ii}$, where α = arbitrary constant and $\hat{C}_{ii}$ are fixed numbers. Let $B_{kk} = \{\alpha\hat{C}_{11}/2\mu_1, \ldots, \alpha\hat{C}_{nn}/2\mu_n\}$.

(2) Select α so that

$$\alpha = 2\mu_k K/\epsilon^2 \hat{C}_{kk} \tag{44}$$

(3) Using the value of K in (18) and the above C matrix, calculate K' by (39) and (40);

(4) Finally obtain ν by (11) where

$$M = \max \{|2B_{mm}\epsilon|, z\} \quad , \tag{45}$$

where

$$B_{mm} = \max \{B_{11}, \ldots, B_{nn}\} \tag{46a}$$

$$z = \max (\hat{\varphi}(\sigma_M), -\hat{\varphi}(-\sigma_M)) \tag{46b}$$

An extension of this theory as well as its application to some examples will be submitted for publication in the near future.

Finally, we would like to point out that Goldwyn, Chao and Chang [4] have recently used an interesting Liapunov function formulation to establish bounded input-bounded output stability of forced direct control systems which satisfy a Popov-type condition. While we also use a Liapunov function formulation, the problem treated here as well as our approach in estimating the bounds are clearly different from those in [4].

7. REFERENCES

[1] J. P. LaSalle and S. Lefschetz, Stability by Liapunov's Direct Method with Applications (book), Academic Press, New York, 1961, pp. 72-73.

[2] op. cit., section 20.

[3] S. Lefschetz, Stability of Nonlinear Control Systems (book), Academic Press, New York, 1965, pp. 24-26.

[4] R. M. Goldwyn, K. S. Chao, and C. Y. Chang, Bounded Input-Bounded Output Stability of Systems, Rice University Technical Report EE-66 No. 9, May 1966; submitted for publication.

A STUDY OF SYSTEM BEHAVIOR FOR NONNORMAL SYSTEMS NEAR AND AT INFINITY OF THE PHASE PLANE

WAYNE H. CHEN
Department of Electrical Engineering
University of Florida, Gainesville, Florida

ABSTRACT

For nonlinear systems $dx/dt = P(x,y)$, $dy/dt = Q(x,y)$ where $P(x,y)$ and $Q(x,y)$ are polynomials of degrees p and q respectively, a study of the behavior of normal systems ($p = q$) at infinity of the phase plane has been reported by the author (1965 Allerton Proceedings, pp. 546-562); and this procedure, however, fails for nonnormal systems ($p \neq q$).

This paper presents a method of studying, for nonnormal systems, the system behavior at infinity by first 1) introducing an artificial term $\epsilon^m f_h(x,y)$ to obtain a "modified" normal system; 2) obtaining its "complete portrait"; and 3) letting $\epsilon \to 0$ to obtain the "complete portrait" of the nonnormal system. In this process, we note in general that several elementary signular points will move closer and closer until they coalesce into a "non-elementary singular point" as $\epsilon \to 0$; and that some of these non-elementary singular points may be located on the idealized boundary E at infinity of the phase plane.

In the example of $dx/dt = y$, $dy/dt = -2x + 3y + x^2$, we first obtain a "modified" normal system $dx/dt = y + \epsilon^3 x^2$, $dy/dt = -2x + 3y + x^2$. Following the suggested process, a "saddle point" and a "focus" (both in the finite region of the phase plane for $\epsilon \neq 0$) and a "node" on E at infinity are found to coalesce, for $\epsilon \to 0$, into a "non-elementary singular point" on the idealized boundary E at infinity of the phase plane.

The method introduced in the paper also enables us to obtain "complete qualitative solutions" of a class of nonlinear differential equations, e.g., $(d^2x/dt^2) = -2x + (dx/dt) + x^2$, and $(d^2x/dt^2) = 3x + 2(dx/dt) + x(dx/dt) + 5(dx/dt)^2 + 4(dx/dt)^3$, which are readily reducible to "nonnormal systems" by introducing $y = dx/dt$.

NETWORK SYNTHESIS USING LOSSY REACTANCES*

JONNY ANDERSEN and HARRY B. LEE
Department of Electrical Engineering and Research Laboratory of
Electronics, Massachusetts Institute of Technology, Cambridge, Mass.

Abstract

A unified synthesis theory is developed for RLCT networks in which the reactances exhibit losses of the semi-uniform type (i.e., all inductors have one Q; all capacitors have another Q). The approach used is to transform networks that contain lossy reactances into networks that contain more nearly ideal reactances. Conventional RLCT synthesis techniques are then applied. The approach proves remarkably successful. Indeed, the final theory is almost as comprehensive as that presently available for RLCT networks.

The paper treats 1) realizability conditions, 2) the Brune procedure, 3) the Darlington (driving point) procedure, 4) the Bott-Duffin procedure, and finally 5) the insertion loss theory.

Manuscript

To appear in the IEEE Transactions on Circuit Theory; September 1966.

---

* This work was supported in part by the Joint Services Electronics Program under Contract DA 36-039-AMC-03200(E).

PRIMITIVE MATRICES FOR DOUBLY TERMINATED LADDER NETWORKS

THOMAS G. MARSHALL: JR.
Radio Corporation of America
Princeton, N. J.

ABSTRACT

A primitive matrix is defined for a doubly terminated ladder network so that the reflection zeros are either equal to the natural frequencies or are their right half-plane mirror images. It is found, for arbitrary distributions of reflections zeros between the right and left half-planes of the frequency variable, that a primitive matrix contains two submatrices, each of which corresponds to a singly terminated network. An algorithm is developed for the synthesis of the matrix of a singly terminated ladder network by a sequence of similarity transformations. The algorithm provides a method for obtaining the continued fraction expansion of the ratio of an nth order polynomial in factored form (and the even or odd part of this polynomial, as n is odd or even respectively) without requiring the factors to be multiplied together or the denominator to be explicitly determined.

INTRODUCTION

The synthesis of a doubly terminated ladder network, which is lossless except for resistances in the terminal branches, can be accomplished by transforming a singly terminated network into an equivalent doubly terminated network by successive perturbations of an associated tridiagonal matrix.[1] A matrix-perturbation algorithm has been developed for the synthesis of RLC ladder networks, such as those in Figure 1; the ρ_i in the series and shunt branches are resistance and conductance values, respectively, and the ϵ_i are, respectively, inductance and capacitance values. The algorithm applies to doubly terminated networks which are the special cases of the illustrated network in which the internal dissipation constants, $\delta_i = \rho_i / \epsilon_i$, are equal to zero (i.e., $\delta_i = 0$, $1 < i < n$). In this perturbation synthesis method, the root-finding of insertion-loss synthesis is replaced by another iterative operation: the generation of a set of equivalent ladder networks of gradually

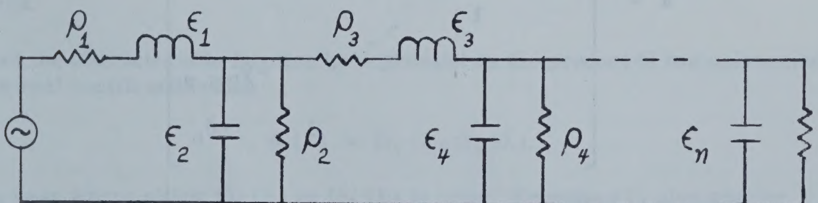


Figure 1. Doubly Terminated RLC Ladder Network

changing termination ratio, ρ_n/ρ_1. During synthesis, the gradual change of element values can be observed, and the intermediate networks can themselves be used if desired. High numerical accuracy was observed, as might be expected of a perturbation algorithm. This accuracy is believed to be attributed, in part, to the formulation of the algorithm so that polynomials are not formed. These desirable features appeared to warrant developing the procedure further to include networks with reflection zeros in both half-planes, which can not be directly obtained by perturbation of singly terminated networks. The perturbation algorithm of Reference 1 will still be applicable in the extended synthesis procedure, so the following discussion will be restricted to the definition and synthesis of primitive matrices that play a role analogous to that of the singly terminated network. The term "primitive" is suggested because the primitive matrices are initial matrices from which the matrices of the doubly terminated networks are derived. The latter matrices could be termed "derivative matrices" by analogy, although this term will not be used herein.

The desirable feature of the singly terminated network retained in the analogous concept of the primitive matrix is that the reflection zeros are the same as the natural frequencies of the network or are their right half-plane mirror images. The reflection zeros are therefore obtainable without the root-finding that is usually required in the insertion-loss synthesis procedure. Primitive networks other than the singly terminated networks do not exist because certain of the elements are infinite in value. However, the matrix introduced in Reference 1 to describe ladder networks does exist for all the reflection zero distributions, and this matrix formulation of ladder network synthesis will be employed.

The discussion will be primarily restricted to ladder networks, such as that in Figure 1, which have no finite zeros of transmission. An algorithm for the synthesis of primitive matrices for this class of ladder networks will be presented. A generalization of the concept of primitive matrix will be mentioned, however, which is applicable to more complex ladder networks. In the next section, a primitive matrix will be defined and shown to be unique.

DEFINITION OF A PRIMITIVE MATRIX

In matrix A below, where $\chi_{i,i+1} = 1/\epsilon_i\, \epsilon_{i+1}$ denotes a coupling constant,

$$A = \begin{bmatrix} -\delta_1 & -\chi_{12} & & & \\ 1 & -\delta_2 & -\chi_{23} & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & -\chi_{n-1,n} \\ & & & 1 & -\delta_n \end{bmatrix} \tag{1}$$

the natural frequencies of the network are the eigenvalues.[1,2] Let the characteristic function of matrix A, $|\lambda I - A|$, be denoted $\Phi(\lambda) \equiv \Phi(\lambda; +\delta_1, +\delta_n) \equiv \Phi(++)$. Networks, which are the same as that in Figure 1 except that the sign of ρ_1 and/or ρ_n is changed, have characteristic functions denoted as $\Phi(-+)$, $\Phi(+-)$, and $\Phi(--)$.

The reflection coefficients, $s_{ii}(\lambda)$, are expressible as

$$\pm s_{11}(\lambda) = \frac{\Phi(-+)}{\Phi(++)}, \quad \pm s_{nn}(\lambda) = \frac{\Phi(+-)}{\Phi(++)}, \tag{2}$$

where the plus sign is used when branch 1 or n, as appropriate, is series.

Assuming now that $\delta_i = 0$, $1 < i < n$, the following relations will hold:

$$\Phi(-\lambda; -+) = (-1)^n \Phi(\lambda; +-), \quad \Phi(-\lambda; ++) = (-1)^n \Phi(\lambda; --). \tag{3}$$

These can be established by observing the effect of changing the signs of all the main diagonal entries in the determinants which are the left hand members of Equation (3). The known relation of insertion-loss theory, $s_{11}(-\lambda) = -s_{nn}(\lambda)$, follows from the first relation.

From insertion-loss theory and Equations (2) and (3)

$$s_{n1}(\lambda) s_{n1}(-\lambda) + s_{11}(\lambda) s_{11}(-\lambda) = \frac{\zeta}{(-1)^n \Phi(++) \Phi(--)} + \frac{\Phi(-+)\Phi(+-)}{\Phi(++)\Phi(--)} = 1, \tag{4}$$

where

$$\zeta = \frac{4\rho_1 \rho_n}{\gamma^2}, \text{ and}$$

$$\gamma = \epsilon_1, \epsilon_2 \ldots, \epsilon_n. \tag{5}$$

The relation

$$(-1)^n \Phi(++) \Phi(--) = \zeta + (-1)^n \Phi(-+) \Phi(+-) \tag{6}$$

follows from Equation (4). This relation is an identity, not dependent upon ρ_1 and ρ_n being positive as often assumed in establishing Equation (4), and it can be verified by multiplying Equation (6) by γ^2 and employing the identity in simple continuants, $K(1,n)K(2,n-1) - K(1,n-1)K(2,n) = (-1)^n$. [2,3] This latter identity is of fundamental importance in the theory of continued fractions where it is referred to as the determinantal formula[4], so it is of interest to note that Equation (6), which is a particular case of Equation (4), can be used to establish the determinantal formula.

Let the characteristic function be expressed as the product of two polynomials in λ with real coefficients; then

$$\Phi(\lambda) = \Theta_1(\lambda)\Theta_2(\lambda). \tag{7}$$

In the case where either $\Theta_1(\lambda)$ or $\Theta_2(\lambda)$ is unity, Equation (7) also applies.

Let $\Theta_i(+) = \Theta_i(\lambda)$ and let $\Theta_i(-)$ be the polynomial that results when the real parts of the roots of $\Theta_i(\lambda)$ are given the opposite sign. To avoid ambiguity, it will be assumed that all roots of $\Phi(\lambda)$ have nonzero real parts.

Definition: A primitive matrix is a matrix, A, with $\delta_i = 0$, $1 < i < n$, for which

$$\Phi(-+) = \Theta_1(-)\,\Theta_2(+). \tag{8}$$

This definition makes it trivial to determine $\Phi(-+)$, thereby making the determination of a primitive matrix much simpler than the matrix of an arbitrary doubly terminated network. It follows from Equation (3) that

$$\Phi(+-) = \Theta_1(+)\,\Theta_2(-) \tag{9}$$

and therefore

$$\Phi(-+)\,\Phi(+-) = \Phi(++)\,\Phi(--). \tag{10}$$

The reflection coefficients become

$$\pm s_{11}(\lambda) = \frac{\Theta_1(-)}{\Theta_1(+)}, \quad \pm s_{nn}(\lambda) = \frac{\Theta_2(-)}{\Theta_2(+)}, \tag{11}$$

and therefore have unity magnitude for $\lambda = j\omega$. They are seen to be of reduced order. From Equations (6) and (10), it is seen that primitive matrices are the special case in which

$$\zeta = 0. \tag{12}$$

From Equations (12) and (5), it is apparent that some combination of $\rho_1 = 0$, $\rho_n = 0$, and $\epsilon_i \to \infty$ must occur. Alternately, at least one of the matrix entries δ_1, δ_n, or $\chi_{1,i+1}$ must equal zero. It will now be shown that only one such matrix entry is zero, subject to the assumption that all roots of $\Phi(\lambda)$ have non-zero real parts. Note that if both δ_1 and δ_n were equal to zero then A would have all the main diagonal entries equal to zero and, therefore, all roots of $\Phi(\lambda)$ would lie on the $j\omega$ axis, contrary to assumption. Observe that when a $\chi_{i,i+1}$, is zero, A is block lower triangular as illustrated in Equation (13) and $\Phi(\lambda) = |\lambda I - A_1|\,|\lambda I - A_2|$.

$$A = \left[\begin{array}{c:c} A_1 & \\ \hdashline 1 & A_2 \end{array}\right] \tag{13}$$

Any combination of two or more of the entries δ_1, δ_n, or $\chi_{i,i+1}$ equal to zero implies that $\Phi(\lambda)$ has a factor whose roots lie on the $j\omega$ axis, contrary to assumption. Therefore we have the following theorem.

Theorem: A primitive matrix has one and only one of the entries δ_1, δ_n, or $\chi_{i,i+1}$ equal to zero if $\Phi(\lambda)$ has no roots on the $j\omega$ axis.

In the following discussion, it will be assumed that all the roots of $\Phi(\lambda)$ are in the left-half plane, thereby satisfying the restriction of the theorem.

The case where either δ_1 or δ_n equals zero corresponds to the well known case of singly terminated networks, so attention will now be directed to the case where one of the entries $\chi_{i,i+1}$ is equal to zero. A will then have the form given in Equation (13) where A_1 and A_2 are tridiagonal matrices with the main diagonal entries equal to zero except for δ_1 and δ_n respectively.

Observe that the characteristic equation of A_1 is a factor of $\Phi(\lambda)$; that its roots change sign when $\delta_1 \to -\delta_1$; and that no roots of the characteristic equation of A_2 have this property. Therefore

$$|\lambda I - A_1| = \Theta_1(\lambda), \tag{14}$$

and therefore,

$$|\lambda I - A_2| = \Theta_2(\lambda). \tag{15}$$

If the order of $\Theta_1(\lambda)$ is n_1, then $\chi_{n_1, n_1+1} = 0$, $1 < n_1 < n$. (When $\delta_1 = 0$, $n_1 = 0$, and when $\delta_n = 0$, $n_1 = n$.) It is well known that the matrices A_1 and A_2 are uniquely determined by $\Theta_1(\lambda)$ and $\Theta_2(\lambda)$ — also that δ_1 or δ_n, as appropriate, and all $\chi_{i,i+1}$ are positive if all the roots of $\Theta_i(\lambda)$ are in the left half plane; therefore, the primitive matrix is uniquely defined, and if all roots of $\Phi(\lambda)$ are in the left-half plane, δ_1, δ_n, and all $\chi_{i,i+1}$ will be non-negative. (Both the theorem and $\chi_{n_1,n_1+1} = 0$ can also be established by Euclid's algorithm because neither K (2,n) nor K (1,n-1) are prime relative to K (1,n).)

The vanishing of the entry χ_{n_1,n_1+1} can be given a physical interpretation by letting $\chi_{n_1,n_1+1} = \alpha\chi'_{n_1,n_1+1}$, where χ_{n_1,n_1+1} is some non zero constant and $\alpha \to 0$. If the ϵ_i, $1 \leq i \leq n_1$, are assumed independent of α, then $\epsilon_{n_1+1} = 1/\epsilon_{n_1}\chi_{n_1,n_1+1} = \epsilon'_{n_1+1}/\alpha$, $\epsilon_{n_1+2} = \alpha\epsilon'_{n_1+2}$, etc., which indicates that the immittance level of the ϵ_i for $i > n_1$ is changed by the factor $1/\alpha$. This change in immittance level, when $\alpha = 0$, effectively isolates or uncouples the two halves of the network. This is illustrated in Figure 2, where branch n_1 is assumed to be a capacitor.

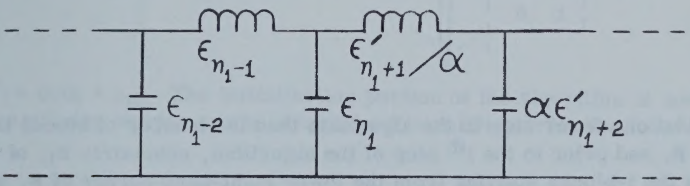


Figure 2. Primitive Network

Primitive networks can be defined in terms of the finite elements $\epsilon_1, \ldots, \epsilon_{n_1}$, $\epsilon'_{n_1+1}, \ldots, \epsilon'_n$ and the parameter α. These elements can be obtained from the reduced reflection coefficients in Equation (11) by normal synthesis techniques. This definition and synthesis procedure for primitive networks could also be extended to more general ladder networks than that of Figure 1. Starting from Equation (11), two singly terminated networks can be synthesized independently. The parameter α will be the impedance conversion factor that must be introduced because the two parts of the network cannot have independently specified impedance levels.

AN ALGORITHM FOR THE SYNTHESIS OF PRIMITIVE MATRICES

The synthesis of primitive matrices is accomplished by the separate synthesis of A_1 and A_2 and juxtaposing them as in Equation (13). It will suffice to treat only the case of a matrix A_1 where δ_1 is the only non-zero δ_i, because the order of the entries of A_1 can be reversed to give a matrix of the form of A_2 without changing the characteristic equation. The synthesis of a matrix A either by a continued fraction expansion or by Lanczos' algorithm has been described.[2] Both of these methods were shown to be applicable to matrices of a general RLC network of the form in Figure 1, and included a matrix such as A_1 as a special case. Several distinct operations are required in these procedures.

An algorithm applicable only to a matrix of the form A_1 will now be presented; this algorithm has only one basic operation which is applied repetitively to accomplish synthesis. The advantage of Lanczos' algorithm over the continued fraction expansion (not having to multiply the factors of $\Theta_1(\lambda)$ together to form a polynomial) is retained, but, in addition, the matrix function, K (2,n), required in Lanczos' algorithm, is no longer required.

A matrix, B, in tridiagonal normal form, such as shown below, which places the linear and quadratic factors of $\Theta_1(\lambda)$ in evidence, has been shown to be similar to any tridiagonal matrix of the form of A_1 which has the same characteristic function and will be considered as the matrix with which the synthesis starts. The order of the factors will not be specified.

$$B = \left[\begin{array}{c|cc|cc|c} -\alpha_1 & & & & & \\ \hline 1 & -\alpha_2 & -\beta_2 & & & \\ & 1 & 0 & & & \\ \hline & & 1 & -\alpha_4 & -\beta_4 & \\ & & & 1 & 0 & \\ \hline & & & & & \ddots \end{array}\right] \sim A_1 \tag{16}$$

There will be one fewer step in the algorithm than the number of blocks in the initial matrix B, and prior to the i^{th} step of the algorithm, submatrix B_i, of the same order as the i-blocks starting from the lower right-hand corner of B, will have been placed in the desired form with only its first main diagonal entry non zero.

The algorithm will first be established for the case where the $(i+1)^{th}$ factor is linear. The objective will be to perform the following transformation:

$$X^{-1} \left[\begin{array}{c|c} -\alpha_i & \\ \hline & B_i \end{array}\right] X = B_{i+1} \tag{17}$$

where both B_i and B_{i+1} are of the desired form. A method of transforming an arbitrary square matrix to the desired form, subject to certain conditions, is known[5] and will be employed here where the conditions are satisfied and where the method is particularly convenient due to the special form of the matrix to be transformed.

Define an elementary transformation matrix to be a matrix, X, which has all plus 1's on the main diagonal and only one other non zero entry, x_{ij}, where $i < j$. The inverse matrix X^{-1} is of the same form with $-x_{ij}$ replacing x_{ij}. The first transformation,

$$\left[\begin{array}{c|cccc} -\alpha_1 & & & & \\ \hline 1 & -\alpha_2 & -\beta_2 & & \\ & 1 & 0 & -\beta_3 & \\ & & & \ddots & \end{array}\right] \sim \begin{bmatrix} -\alpha_1' & -\alpha_1\alpha_2 & \alpha_2\beta_2 & & \\ 1 & 0 & -\beta_2 & & \\ & 1 & 0 & -\beta_3 & \\ & & & \ddots & \end{bmatrix} \tag{18}$$

in which $x_{12} = \alpha_2$ and $\alpha'_1 = \alpha_1 + \alpha_2$, places the main diagonal in the proper form. For convenience in presenting the algorithm, the subscripts have been renumbered starting from unity in Equation (18). The second transformation, in which $x_{13} = \alpha_2\,\beta_2/\alpha'_1$, yields

$$\begin{bmatrix} -\alpha'_1 & -\beta'_1 & 0 & \beta_3\, x_{13} & \\ 1 & 0 & -\beta'_2 \pm x_{13} & & \\ & 1 & 0 & -\beta_3 & \\ & & & \ddots & \end{bmatrix} \tag{19}$$

where $\beta'_1 = \alpha_1\alpha_2 + x_{13}$. The initialization portion of the algorithm is now completed, and the remaining entries can be determined iteratively from the following expressions,

$$\begin{aligned} x_{i,\,i+2} &= \frac{\beta_{i+1}\, x_{i-1,\; i+1}}{\beta'_{i-1}} \\ \beta'_1 &= \beta_i - x_{i-1,\; i+1} + x_{i,\,i+2} \end{aligned} \tag{20}$$

for $2 \leq i \leq n-1$ with the definition $\beta_n \equiv 0$. The primed values are the desired entries of B_{i+1}. Both B_i and B_{i+1} have the general form of A_1 and therefore $\alpha_1, \alpha'_1, \beta_1$, and β'_i are greater than zero. The conditions for using the algorithm are that α'_1, and β'_{i-1}, $2 \leq i \leq n-1$, be nonzero, and they are satisfied (by prior assumptions).

The case where the $(i+1)^{\text{th}}$ factor is quadratic will now be considered. Assuming B_i to be of the desired form, the objective is to perform the indicated transformation,

$$\left[\begin{array}{cc|c} -\alpha_1 & -\beta_1 & \\ 1 & 0 & \\ \hdashline & & B_i \end{array}\right] \sim B_{i+1} , \tag{21}$$

so that B_{i+1} is also of the correct form. A sequence of the following four transformations,

$$x_{23} = \alpha_3, \; x_{14} = -\alpha_3\beta_3, \; x_{12} = \alpha_3, \text{ and } x_{13} = \alpha_3(\beta_3 - \beta_1)/\alpha'_1, \tag{22}$$

where $\alpha'_1 = \alpha_1 + \alpha_3$, yields the following matrix,

$$\begin{bmatrix} -\alpha'_1 & -\beta'_1 & 0 & \beta_3 x_{13} + \alpha_1\alpha_3\beta_3 & -\alpha_3\beta_3\beta_4 \\ 1 & 0 & x_{13} & & \\ & 1 & 0 & -\beta_3 & \\ & & 1 & 0 \; \ddots & -\beta_4 \end{bmatrix} , \tag{23}$$

with $\beta'_1 = \beta_1 + \alpha_1\alpha_3 + x_{13}$. Two more transformations, $x_{15} = -\alpha_3\beta_3\beta_4/\alpha'_1$ and $x_{24} = \left(\beta_3 x_{13} + \alpha_1\alpha_3\beta_3 - x_{15}\right)/\beta'_1$, complete the initialization and yield the following matrix,

$$\begin{bmatrix} -\alpha'_1 & -\beta'_1 & 0 & 0 & 0 & \beta_5 x_{15} \\ 1 & 0 & -\beta'_2 & 0 & x_{15} + \beta_4 x_{24} & \\ & 1 & 0 & -\beta_3 + x_{24} & & \\ & & 1 & 0 & -\beta_4 & \\ & & & & \ddots & \end{bmatrix} \tag{24}$$

where $\beta'_2 = -x_{13} + x_{24}$.

The remaining entries can be calculated iteratively from the following expressions,

$$\begin{aligned} x_{i-1,\,i+3} &= \frac{\beta_{i+2}\, x_{i-2,\,i+2}}{\beta'_{i-2}} \\ x_{i,\,i+2} &= \frac{\beta_{i+1}\, x_{i-1,\,i+1} + x_{i-2,\,i+2} - x_{i-1,\,i+3}}{\beta'_{i-1}} \\ \beta'_i &= \beta_i - x_{i-1,\,i+1} + x_{i,\,i+2} , \end{aligned} \tag{25}$$

for $3 \leq i \leq n-1$ with the definition $\beta_n = \beta_{n+1} \equiv 0$. Note that each iteration nullifies the two entries which are not in the three center diagonals, and, except for the last two iterations, two more appear, but with both row and column subscripts advanced by one. For example, after one iteration, the matrix Equation (24) will have been transformed to one in which entries 16 and 25, appearing in Equation (24), are zero; and a new entry 27, which is $\beta_6 x_{26}$, and a new entry 36, which is $x_{26} + \beta_5 x_{35}$, appear.

It is possible to start the iterations Equation (25) with $i = 2$, i.e. with the matrix of Equation (23), if the following definitions are made,

$$\beta'_0 \equiv -\alpha_1 \alpha'_1, \; x_{04} \equiv \alpha_1 \alpha_3 \beta_3. \tag{26}$$

A further convenience is possible in that this algorithm for a quadratic factor can be used for a linear factor, also, by substituting α_2 for α_3 and β_2 for β_3 in Equation (22) and defining $x_{04} \equiv 0$ for the case of a linear factor. All the $x_{i-1,\, i+3}$ of Equations (25) will therefore be zero and Equation (25) reduces to Equation (20).

To synthesize a matrix of the form A_1, the initialization and iteration just described, is repeated m-1 times, if B has m blocks. This synthesis procedure is simpler to program than the continued fraction or Lanczos' algorithm because there is only one basic step. A 10th order Butterworth ladder network was synthesized by the algorithm using 8-digit accuracy, and the element values were found to have 6-digit accuracy. The perturbation algorithm could improve upon this to give 7-digit accuracy. By contrast, a continued fraction expansion gave 4- to 5-digit accuracy. The number of multiplications and the number of divisions in the algorithm both approach $n^2/2$ for large n.

CONCLUSION

A primitive matrix was defined and shown to be unique, subject to the restriction that the specified natural frequencies have nonzero real parts. The primitive matrix can be used to initiate the synthesis of a doubly terminated network by a known perturbation algorithm. An algorithm was presented for synthesizing the component matrices in a primitive matrix. These component matrices correspond to singly terminated networks, so the algorithm can be used to synthesize this class of network.

BIBLIOGRAPHY

[1] Marshall, T.G., "Discussion of 'Computer solution of the network realization problem,' by D.A. Calahan," *Proc. 2nd Allerton Conference on Circuit and System Theory*, Univ. of Ill., (Sept., 1964) pp. 175-200.

[2] Marshall, T.G., "Ladder network synthesis by similarity transformation," *Proc. 3rd Allerton Conference on Circuit and System Theory*, U of Ill. (Oct. 1965) p. 200.

[3] Muir, T., rev. ed. by W.H. Metzler, *A Treatise of the Theory of Determinants*, Dover Publications, Inc.; New York, N.Y. (1960).

[4] Wall, H.S., *Analytic Theory of Continued Fractions*, D. Van Nostrand Co., Princeton, N.J. (1948).

[5] Schwartz, H.R., "Ein Verfahren zur Stabilitatsfrage bei Matrizen-Eigenwertproblemen," *Z. angew. Math. u Phys.*, vol. 7 (1956) pp. 473-500.

REALIZATION OF THE EQUILIBRIUM AND RESPONSE IMMITTANCES OF TRANSFORMERLESS RLC NETWORKS

J. E. OLIVARES, JR.
Hughes Aircraft Company
Culver City, California

A recent paper$^{(1)}$ by the author has presented existence theorems for realizability of the equilibrium immittance X and its inverse the solution or response immittance W of a resistance (R) network with (n + 1) nodes and b branches. This paper will extend the existence theorems to RLC networks without mutual inductances and show a method for realizing the response immittance W(s) for an RLC network with known topology. This paper thus provides new results which establish a test for realizability of the equilibrium and response immittances of transformerless RLC networks. This test is applicable to an (n + 1) node RLC network with a complete graph referred to a star-tree or any other specified tree (e.g., linear tree) or, in fact, any network with known topology; that is, a network whose tie-set and cut-set matrices are known.

RLC Network Analysis

Consider a general (n + 1) node RLC network with b branches (Figure 1). Each generalized network branch is assumed to consist of a single kind of passive element - resistance (R), capacitance (C) and inductance (L) - with no mutual inductances and with a series voltage source and/or parallel current source.

Thus, depending upon the type of branch, the individual branch immittances may be represented as

$$f_i(s) = k_\infty^{(i)} s, \quad \frac{k_o^{(i)}}{s}, \quad \text{or } k^{(i)}$$

and the branch impedance matrix R(s) and branch admittance matrix G(s) are diagonal matrices with elements $f_i(s)$.

The node-pair admittance matrix $Y_n(s)$ referred to an arbitrary network tree relates the node-pair voltages $\underline{e}$ to the current sources $\underline{j}_s$ by the equation

$$Y_n(s)\, \underline{e} = \underline{j}_s \tag{1}$$

where $Y_n(s) = A_n G(s) A_n^T$ (2)

A_n = cut-set matrix of order n x b

A_n^T = matrix transpose of A_n

and $G(s)$ = diagonal branch admittance matrix of order b

$= \left[g_i(s)\, \delta_{ij}\right]$

where δ_{ij} = Kronecker's delta.

For a connected network, the matrix $Y_n(s)$ is of order n and has rank n, i.e., it is nonsingular.

The loop impedance matrix $Z(s)$ relates the loop currents $\underline{i}$ to the voltage sources $\underline{v}_s$ by

$$Z_{\ell}\underline{i} = \underline{v}_s \tag{3}$$

where $$Z_{\ell}(s) = B_{\ell}R(s)\, B_{\ell}^{T} \tag{4}$$

B_{ℓ} = tie-set matrix of order $(b - n) \times b$

$R(s)$ = diagonal branch impedance matrix of order b

that is $R(s) = G^{-1}(s)$

The matrix $Z_{\ell}(s)$ in this case has order and rank $\ell = b - n$. Furthermore, the sum of the ranks of the node-pair admittance and loop impedance matrix is exactly b, the number of network branches.

Assume that the generalized network branches have a branch current vector $\underline{j}$ and branch voltage vector $\underline{v}$, which by a suitable branch numbering system may be partitioned into link branches and tree branches as follows:

$$\underline{j} = \left[\begin{array}{c} \underline{j}_1 \\ \hline \underline{j}_2 \end{array}\right] \qquad \underline{v} = \left[\begin{array}{c} \underline{v}_1 \\ \hline \underline{v}_2 \end{array}\right]$$

where $\underline{j}_1$ and $\underline{v}_1$ correspond to the link branches and $\underline{j}_2$ and $\underline{v}_2$ correspond to the tree branches.

Furthermore, let the loop currents $\underline{i}$ and node-pair voltage drops $\underline{e}$ be related to the link currents and tree branch voltages by

$$\underline{i} = \underline{j}_1 \text{ and } \underline{e} = \underline{v}_2$$

Then, the cut set matrix and tie set matrix are given by

$$A_n = \left[\begin{array}{c|c} P & I_n \end{array}\right] \quad \text{and} \quad B_{\ell} = \left[\begin{array}{c|c} I_{\ell} & -P^{T} \end{array}\right] \tag{5}$$

where I_k = identity matrix of order k

and P = unimodular matrix of order $n \times \ell$

Then, partitioning the branch admittance matrix $G(s)$ and branch impedance matrix $R(s)$ into link and tree branches results in

$$G(s) = \left[\begin{array}{c|c} G_{\ell}(s) & 0 \\ \hline 0 & G_t(s) \end{array}\right] \text{ and } R(s) = \left[\begin{array}{c|c} R_{\ell}(s) & 0 \\ \hline 0 & R_t(s) \end{array}\right] \tag{6}$$

where $G_{\ell}(s)$ and $R_{\ell}(s)$ are link branch immittances and $G_t(s)$ and $R_t(s)$ are tree branch immittances.

The, substituting Equations (5) and (6) in (2) and (4) gives

$$Y_n(s) = G_t(s) + PG_{\ell}(s)\, P^{T} \tag{7}$$

and $$Z_\ell(s) = R_\ell(s) + P^T R_t(s) P \tag{8}$$

For a connected network with passive elements, the matrices $Z_\ell(s)$ and $Y_n(s)$ are each positive definite. In order to obtain a one-to-one correspondence between network branches and equilibrium equations both the node-pair admittance and loop impedance matrix must be specified to determine network realizability.

If we define the equilibrium immittance matrix $X(s)$ as the direct sum of $Z_\ell(s)$ and $Y_n(s)$; that is,

$$X(s) = \left[\begin{array}{c|c} Z_\ell(s) & 0 \\ \hline 0 & Y_n(s) \end{array}\right] = Z_\ell(s) \oplus Y_n(s) \tag{9}$$

then $X(s)$ is a positive definite matrix and can be written as the congruence

$$X(s) = \left[\begin{array}{c|c} R_\ell(s) & -P^T \\ \hline P & G_t(s) \end{array}\right] \left[\begin{array}{c|c} G_\ell(s) & 0 \\ \hline 0 & R_t(s) \end{array}\right] \left[\begin{array}{c|c} R_\ell(s) & P^T \\ \hline -P & G_t(s) \end{array}\right] \tag{10}$$

or $$X(s) = J(s)\, D^{-1}(s)\, J^T(s) \tag{11}$$

where $$J(s) \triangleq \left[\begin{array}{c|c} R_\ell(s) & -P^T \\ \hline P & G_t(s) \end{array}\right]$$

$$D(s) \triangleq \left[\begin{array}{c|c} R_\ell(s) & 0 \\ \hline 0 & G_t(s) \end{array}\right]$$

and $$D^{-1}(s) = \text{matrix inverse of } D(s)$$

$$= \left[\begin{array}{c|c} G_\ell(s) & 0 \\ \hline 0 & R_t(s) \end{array}\right]$$

Furthermore, the response immittance matrix $W(s) = X^{-1}(s)$, the direct sum of $Y_\ell(s)$ and $Z_n(s)$, defined by

$$W(s) = \left[\begin{array}{c|c} Y_\ell(s) & 0 \\ \hline 0 & Z_n(s) \end{array}\right] = Y_\ell(s) \oplus Z_n(s) \tag{12}$$

is given by $$W(s) = K^T(s)\, D(s)\, K(s) \tag{13}$$

where $K(s) \triangleq J^{-1}(s)$

$$= \left[\begin{array}{c|c} (R_\ell(s) + P^T R_t(s) P)^{-1} & G_\ell(s) P^T (G_t(s) + P G_\ell(s) P^T)^{-1} \\ \hline -R_t(s) P (R_\ell(s) + P^T R_t(s) P)^{-1} & (G_t(s) + P G_\ell(s) P^T)^{-1} \end{array}\right] \tag{14}$$

$$K(s) = W(s) + D^{-1}(s)Q^T W(s) = D^{-1}(s)J^T(s)W(s)$$

Furthermore, the matrix W(s) is congruent to D(s) since

$$J^T(s)\, W(s)\, J(s) = D(s) \tag{15}$$

Note that the matrix J(s) can be written as the sum of the diagonal matrix D(s) and the skew-symmetric matrix Q of the form

$$J(s) = D(s) + Q \tag{16}$$

where

$$Q = \left[\begin{array}{c|c} 0 & -P^T \\ \hline P & 0 \end{array}\right] = -Q^T$$

Hence $$X(s) = (D(s) + Q)D^{-1}(s)(D(s) + Q^T) \tag{17}$$

It can be shown(2) that

$$X(s) = D(s) + QD^{-1}(s)Q^T \tag{18}$$

On the other hand, since

$$(D(s) + Q^T)\, W(s)\, (D(s) + Q) = D(s) \tag{19}$$

It can be shown that

$$D(s)W(s)D(s) + Q^T\, W(s)\, Q = D(s) \tag{20}$$

This leads to the formal statement of a realizability theorem for both the equilibrium and response immittances of a passive transformerless RLC network.

Theorem I:

A symmetric matrix X(s) of order b, the direct sum of the loop impedance matrix $Z_\ell(s)$ and node-pair admittance matrix $Y_n(s)$ defined by

$$X(s) = \left[\begin{array}{c|c} Z_\ell(s) & 0 \\ \hline 0 & Y_n(s) \end{array}\right] = Z_\ell(s) \oplus Y_n(s)$$

where $Z_\ell(s)$ and $Y_n(s)$ are positive definite matrices of order and rank $\ell = b - n$ and n respectively is the equilibrium immittance matrix of a connected (n + 1) node RLC network with b passive branches and no mutual inductances if and only if

(1) X(s) can be decomposed in the congruence

$$X(s) = J(s)\, D^{-1}(s)\, J^T(s)$$

(2) The inverse branch immittance matrix $D^{-1}(s) = \left[d_i^{-1}(s)\ \delta_{ij}\right]$

$$= \left[\begin{array}{c|c} \mathcal{G}_\ell(s) & 0 \\ \hline 0 & R_t(s) \end{array}\right]$$

is a diagonal matrix of order b containing elements

$$d_i^{-1}(s) = k_\infty^{(i)} s,\ \frac{k_o^{(i)}}{s},\ \text{or } k^{(i)}$$

where $k_\infty^{(i)}$, $k_o^{(i)}$ and $k^{(i)}$ are all positive.

(3) The matrix J(s) is a nonsingular matrix of order b that has the form

$$J(s) = \left[\begin{array}{c|c} R_\ell(s) & -P^T \\ \hline P & G_t(s) \end{array}\right] = D(s) + Q$$

(4) The unimodular matrix Q is skew symmetric of the form

$$Q = \left[\begin{array}{c|c} 0 & -P^T \\ \hline P & 0 \end{array}\right]$$

(5) The matrix P is realizable unimodular, that is,

$A \triangleq \left[P \mid I_n\right]$ is a realizable as a cut-set matrix

or $B \triangleq \left[I_\ell \mid -P^T\right]$ is realizable as a tie-set matrix

Theorem II:

A symmetric matrix W(s) of order b, the direct sum of the loop admittance matrix $Y_\ell(s)$ and the node-pair impedance matrix $Z_n(s)$ defined by

$$W(s) = \left[\begin{array}{c|c} Y_\ell(s) & 0 \\ \hline 0 & Z_n(s) \end{array}\right] = Y_\ell(s) \oplus Z_n(s)$$

where $Y_\ell(s)$ and $Z_n(s)$ are positive definite matrices of order and rank $\ell = b - n$ and n respectively is the response immittance matrix of a connected (n + 1) node RLC network with b passive branches and no mutual inductances if and only if:

(1) There exists a nonsingular matrix J(s) that diagonalizes W(s) by the congruent transformation

$$J^T(s)\, W(s)\, J(s) = D(s)$$

(2) The branch immittance matrix D(s) has the form

$$D(s) = \left[d_i(s)\, \delta_{ij}\right] = \left[\begin{array}{c|c} R_\ell(s) & 0 \\ \hline 0 & G_t(s) \end{array}\right]$$

and is a diagonal matrix of order b containing elements

$$d_i(s) = k_\infty^{(i)} s,\ \frac{k_o^{(i)}}{s},\ \text{or } k^{(i)}$$

where $k_\infty^{(i)}$, $k_o^{(i)}$ and $k^{(i)}$ are all positive.

(3) The matrix J(s) is a nonsingular matrix of order b that has the form

$$J(s) = \left[\begin{array}{c|c} R_{\ell}(s) & -P^T \\ \hline P & G_t(s) \end{array}\right] = D(s) + Q$$

(4) The unimodular matrix Q is skew symmetric of the form

$$Q = \left[\begin{array}{c|c} 0 & -P^T \\ \hline P & 0 \end{array}\right]$$

(5) The matrix P is realizable unimodular; i.e.,

$A \triangleq \left[\begin{array}{c|c} P & I_n \end{array}\right]$ is realizable as a cut-set matrix

or $B \triangleq \left[\begin{array}{c|c} I_{\ell} & -P^T \end{array}\right]$ is realizable as a tie-set matrix

Realization of the Equilibrium Immittance for RLC Networks

An (n + 1) node RLC network with b branches can be realized from the equilibrium immittance X(s) by applying the following procedure; that is,

Let $X(s) = Z_{\ell}(s) + Y_n(s)$

where $Z_{\ell}(s) = Ls + \Omega + \Delta/s$ (21)

where L = loop inductance matrix
Ω = loop resistance matrix
Δ = loop elastance matrix

and $Y_n(s) = Cs + \Sigma + \Gamma/s$ (22)

where C = node-pair capacitance matrix
Σ = node-pair conductance matrix
Γ = node-pair inverse inductance matrix

Cederbaum's (3) procedure can be applied to each of the component matrices L, Ω and Δ of $Z_{\ell}(s)$ and the component matrices C, Σ and Γ of $Y_n(s)$ to determine the non-redundant cut-set and tie-set matrices, i.e.,

$$A'_n = \left[\begin{array}{c|c} P' & I_n \end{array}\right] \quad \text{and} \quad B'_{\ell} = \left[\begin{array}{c|c} I_{\ell} & -P'^T \end{array}\right]$$

As shown previously by the author, (4) this procedure results in the non-redundant column vectors of P and P^T. The redundant column vectors can be found by choosing a compatible set of the non-redundant vectors which satisfy the equation

$$P = (P^T)^T$$

After a consistent set of matrices A_n and B_{ℓ} are found the branch immittance matrices R(s) and G(s) can be determined and the resulting network realized.

Realization of the Response Immittance for RLC Networks

A more interesting problem that is closely related to the transformerless N-port synthesis problem for one, two, or three kinds of passive elements is connected with the realization of the response immittance W(s). It will be recalled that the response immittance represents the dynamical description of the network at the terminals where all node-pairs and loops are accessible. Thus, if a response immittance can be realized directly from realizability conditions based on the response equations rather than the equilibrium equations, this would represent a significant contribution to N-port realizability theory.

This question is considered in this section in which, given the response immittance W(s) and the topology of the network so that the matrix Q can be inferred, a direct realization of the network can be achieved.

From the realizability theorem, a necessary and sufficient condition that the matrix W(s) is the response immittance of a connected (n + 1) node transformerless RLC network with a complete graph is that there exists a matrix J(s) such that

$$J^T(s)\, W(s)\, J(s) = D(s)$$

To determine such a decomposition in general (that is, not knowing either D(s) or Q) still remains an unsolved problem. However, given the topology of the network so that the matrix Q can be inferred leads to a solution of the problem of realizing the network directly from a specification of the response immittance. For example, a method has been devised for decomposing the response immittance of an (n + 1) node RLC network with a complete graph referred to a star-tree. The linear tree or any other specific tree can be treated by applying a congruent (linear-star tree) transformation. The method is also applicable for any other network given the network topology.

The method therefore provides a constructive test for establishing that a matrix W(s) is the star-tree response immittance of an (n + 1) node RLC network with a complete graph. For this case, it has been shown(5) that the unimodular matrix P is given by the recursive representation

$$P \triangleq E_n = \left[\begin{array}{c|c} E_{n-1} & I_{n-1} \\ \hline \underline{0}_{\ell'}^{T} & -\underline{\beta}_{n-1}^{T} \end{array}\right] \text{ for } n = 3, 4, \ldots\ldots\ldots \quad (24)$$

where $E_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

I_{n-1} = identity matrix of order n-1

$\underline{\beta}_{n-1}^{T}$ = (n-1) dimensional row vector of unit elements

$\underline{0}_{\ell'}^{T}$ = $(\ell - n + 1)$ dimensional row vector of null elements

The unimodular matrix P corresponding to an (n + 1) node network with a complete graph for other types of trees can be found by means of a suitable transformation. From this, the matrix Q is formed from

$$Q = \left[\begin{array}{c|c} 0 & -P^T \\ \hline P & 0 \end{array}\right]$$

and it only remains to find D(s).

Given the topological structure of the network, a positive test of the response immittance W(s) is given as follows:

since $$DWD + Q^T WQ = D$$

Let $$Q^T WQ \triangleq T$$

then $$D^{-1}\,T\,D^{-1} + W = D^{-1}$$

This equation is sufficient to determine the matrix D(s) given the matrices W(s) and Q since

$$D^{-1}\,T\,D^{-1} - D^{-1} + W = 0$$

If we let D^{-1} have diagonal elements δ_k

and $T = \begin{bmatrix} t_{ij} \end{bmatrix}$ and $W = \begin{bmatrix} w_{ij} \end{bmatrix}$ and $N = D^{-1}\,T\,D^{-1}$

Then

$$N = \begin{bmatrix} \delta_1 t_{11} \delta_1 & \delta_1 t_{12} \delta_2 & \cdots\cdots & \delta_1 t_{1b} \delta_b \\ \delta_2 t_{21} \delta_1 & \delta_2 t_{22} \delta_2 & \cdots\cdots & \delta_2 t_{2b} \delta_b \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \delta_b t_{b1} \delta_1 & \delta_b t_{b2} \delta_2 & \cdots\cdots & \delta_b t_{bb} \delta_b \end{bmatrix} \tag{25}$$

hence

$$\left[D^{-1}\,TD^{-1} - D^{-1} + W\right]_{kk} = \delta_k\, t_{kk}\, \delta_k - \delta_k + w_{kk}$$

or $$\delta_k^2\, t_{kk} - \delta_k + w_{kk} = 0$$

or $$\delta_k = \frac{1 \pm \sqrt{1 - 4t_{kk}\, w_{kk}}}{2t_{kk}} \qquad k = 1, 2 \ldots b. \tag{26}$$

Thus, the elements δ_k of $D^{-1}(s)$ are defined to within one of two solutions of a quadratic by the diagonal terms alone. It will be shown that one solution provides the branch immittance b_k while the other provides the corresponding driving point immittance across the branch. The ambiguity can be resolved from the off-diagonal elements; that is,

$$\delta_i\, t_{ij}\, \delta_j + w_{ij} = 0$$

or $$\delta_i\, \delta_j = -\frac{w_{ij}}{t_{ij}} \qquad \text{for } t_{ij} \neq 0 \tag{27}$$

A similar test can be applied to a network referred to other trees (for example, a linear tree) by applying the linear tree to star tree transformation(6)

$$W^s = H\, W^{\ell} H^T$$

where $$H = \left[\begin{array}{c:c} S_{\ell} & 0 \\ \hdashline 0 & M_n \end{array}\right]$$

and $\underline{i} = S_{\ell}\, i'$ and $\underline{e} = M_n\, \underline{e}'$

The unprimed variables refer to the star tree and the primed to the linear tree.

Realization of the Star-Tree Response Immittance for an RLC Network

This approach can be employed to realize the star-tree response immittance matrix W(s) for an RLC network with a complete graph. For example, let

$$W(s) \doteq \frac{1}{s^2+s+1} \begin{bmatrix} s & 0 & 0 \\ 0 & s^2+1 & 1 \\ 0 & 1 & s+1 \end{bmatrix}$$

$$= \frac{1}{s^2+s+1} \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} s & 0 & 0 \\ 0 & s^2+1 & 1 \\ 0 & 1 & s+1 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

or $T = \frac{1}{s^2+s+1} \begin{bmatrix} s^2+s & 0 & 0 \\ 0 & s & -s \\ 0 & -s & s \end{bmatrix}$

Since $\delta_k \doteq \frac{1 \pm \sqrt{1 - 4t_{kk}\, w_{kk}}}{2t_{kk}}$

then $\delta_1 = \frac{(s^2+s+1) \pm (s^2-s-1)}{2s\,(s+1)} = \frac{2\,(s+1)}{2\,(s+1)} = \frac{1}{s}$

or $\delta_1' = \frac{2s^2}{2s(s+1)} = \frac{s}{s+1} = \frac{1}{1+1/s}$ (which is not of the allowable form $f_i(s)$)

$$\delta_2' = \frac{(s^2+s+1) \pm (s^2-s+1)}{2s} = \frac{2s}{2s} = 1$$

or $\delta_2' = \frac{2(s^2+1)}{2s} = \frac{s^2+1}{s} = s + \frac{1}{s}$ (which is not allowable)

Finally

$$\delta_3' = \frac{(s^2+s+1) \pm (s^2+s-1)}{2s} = \frac{2}{2s} = \frac{1}{s}$$

or $\delta_3' = \frac{2s(s+1)}{2s} = s+1$ (which is not allowable) hence

$$D^{-1}(s) = \begin{bmatrix} 1/s & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/s \end{bmatrix} = \left[\begin{array}{c:c} G_{\ell}(s) & 0 \\ \hdashline 0 & R_t(s) \end{array}\right]$$

hence the RLC network is as shown in Figure 2.
It is quite evident that this method provides a means to determine the branch immittances which are elements of $D^{-1}(s)$. Also, the alternate solutions of the quadratic equations give the driving point immittance across the branch; that is δ'_1 represents the admittance of the circuit in parallel across branch b_1.

Summary and Conclusions

This paper has shown for a transformerless (n + 1) node RLC network that the equilibrium and response immittances

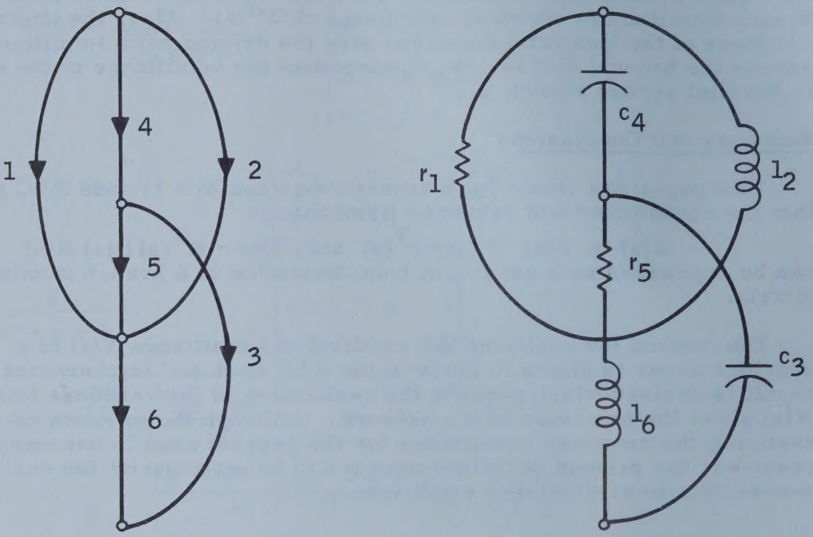
$$X(s) = J(s)\ D^{-1}(s)\, J^T(s) \text{ and } W(s) = K^T(s)\, D(s)\, K(s)$$

can be expressed as a congruent transformation of a branch immittance matrix.

The method for realizing the equilibrium immittance X(s) of a previous paper is shown to apply to the RLC case and furthermore a new result is derived which permits the realization of the response immittance W(s) given the structure of the network. Although the problem of realizing the response immittance for the general case is not completely resolved, the present technique appears to be quite useful for deriving new results in realizability conditions.

References

(1) Olivares, J. E. Jr., "Simultaneous Realization of the Loop Resistance and Node Conductance Matrix by a Topological Approach, "Proceedings of 8th Midwest Symposium on Circuit Theory; Colorado State University, June 14-15, 1965, pp. 24-0 to 24-12.
(2) Olivares, J. E. Jr., op. cit.
(3) Cederbaum, I., "Applications of Matrix Algebra to Network Theory," IRE Transactions CT-6, Special Supplement, 1959, pp 127-137.
(4) Olivares, J. E. Jr., op. cit.
(5) Olivares, J. E. Jr., "Topological Analysis and Synthesis of the Star-Tree Loop Resistance and Node Conductance Matrix" Proceedings of the Midwest Symposium on Circuit Theory, Oklahoma State University, Stillwater, Oklahoma.
(6) Olivares, J. E. Jr., op. cit.


4
1
2
5
3
6
c_4
r_1
l_2
r_5
c_3
l_6

a) Network Graph

b) RLC Network

Figure 1

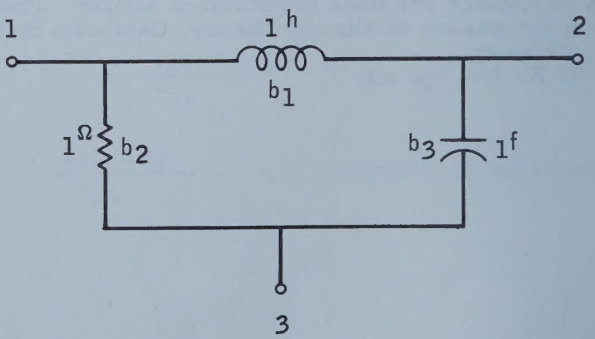

1
1^h
2
b_1
1^Ω
b_2
b_3
1^f
3

Figure 2

NON-SERIES-PARALLEL REALIZATION OF BISYMMETRICAL* TWO-ELEMENT-KIND TWO-PORTS

SAMUEL C. LEE
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

ABSTRACT

A new non-series-parallel synthesis procedure is introduced to synthesize a class of two-element-kind bisymmetrical* network function specifications. Necessary and sufficient conditions for the realizability of the network functions are given. For a set of realizable functions, two non-series-parallel realizations (denoted Form I and Form II) exist. There is a unique realization of Form I, but infinitely many realizations of Form II. A numerical example illustrates the method.

This study uses the results of the realization of symmetrical ladder networks. In this paper, the necessary and sufficient conditions for a set of RC z (or y)-parameters to be realized as a symmetrical (electrically and structurally) canonical ladder network are presented and proven. RC-to-RL and RC-to-LC transformations permit these conditions to be applied to RL and LC cases.

I. INTRODUCTION

Networks which are symmetrical with respect to a pair of ports are useful and often arise in network design. An interesting consequence of the use of an electrically and structurally symmetrical network is the potential economy afforded by the need for a smaller number of different element values and the often smaller range of element values. Furthermore, simplification of both the computation and fabrication processes may result. Thus it is worthwhile to be able to predict when an electrically and structurally symmetrical network may be obtained to realize a specified network function. In this paper, the interest is focused on the realization of non-series-parallel bisymmetrical networks.

Kim [1] and Van Valkenburg [2] have considered the canonical synthesis of a subset of the set of minimum biquadratic and bicubic driving point functions using a non-series-parallel structure. They showed that the non-series-parallel realizations obtained, containing fewer elements than the equivalent Brune realization, are canonical realizations without ideal transformers. Lee and Frisch [3] extended their work to the non-series-parallel realization of minimum network functions of any order. The works cited above all dealt with one-port synthesis.

A non-series-parallel realization of a class of vertically symmetrical RC two-ports (Fig. 1) was recently found [4], with a multiparameter sensitivity (see Sec. VI) always smaller than that of the equivalent series-parallel realization. In this paper, a class of bisymmetrical two-element-kind networks is considered. The results for RC networks can be easily extended to

---

* A two-port network is called bisymmetrical if it possesses both horizontal and vertical symmetry, electrically and structurally.

RL and LC cases. Furthermore, the non-series-parallel synthesis method given in this paper can be easily applied to potentially bisymmetrical non-series-parallel networks.

II. NECESSARY AND SUFFICIENT CONDITIONS FOR THE REALIZABILITY OF SYMMETRICAL LADDER NETWORKS

Cauer [5] first found a general synthesis procedure to realize a set of two-element-kind z-parameters, z_{11}, z_{12} and z_{22}, when the residue condition [ref. 6, p. 311] is satisfied. However, his method in general requires ideal transformers which can only be approximated in practice. The problem of realizing a set of two-element-kind z (or y)-parameters which satisfy the Fialkow-Gerst coefficient conditions [ref. 6, p. 254, and ref. 7] without using mutual inductances is still unsolved. However, an attempt has been made to consider the realization of a set of two-element-kind z (or y)-parameters with $z_{11} = z_{22}$ (or $y_{11} = y_{22}$), i.e., the realization of symmetrical networks [8].

Consider the question: "What are the conditions imposed on a set of two-element-kind z (or y)-parameters with $z_{11} = z_{22}$ (or $y_{11} = y_{22}$) for a symmetrical canonical[†] ladder network realization to exist?" The question is completely answered in this section; the results will be used in our later development.

Theorem: The necessary and sufficient condition for a set of RC z_{11}, z_{12} and z_{22} parameters to be realized as an electrically and structurally symmetrical canonical ladder network as shown in Figs. 2(a) and (b) is that

(a) for N_a, z_{11}, z_{12} and z_{22} can be expressed as

$$z_{11} = z_{22} = \frac{A}{s} + \sum_{i=1}^{n} \frac{A\,|\,\rho_i\,|}{s + \sigma_i} \tag{1}$$

$$z_{12} = \frac{A}{s} + \sum_{i=1}^{n} (-1)^i \frac{A\,|\,\rho_i\,|}{s + \sigma_i} \tag{2}$$

where

$$\rho_i = \lim_{s \to -\sigma_i} \frac{(s+\sigma_i)\left[\prod_{j=1}^{n} \sigma_j\right]}{s \cdot \left[\prod_{j=1}^{n} (s+\sigma_j)\right]} \tag{3}$$

$\sigma_i < \sigma_j$ for $i < j$, and A is a positive real constant;

A two-port network realization is called canonical if it contains the minimum number of circuit elements with respect to a given network function or a set of network functions.

(b) for N_b, z_{11}, z_{12} and z_{22} have the form

$$z_{11} = z_{12} = A + \sum_{i=1}^{n} \frac{A|\rho_i|}{s + \sigma_i} \tag{4}$$

$$z_{12} = A + \sum_{i=1}^{n} (-1)^{n-i+1} \frac{A|\rho_i|}{s + \sigma_i} \tag{5}$$

where

$$\rho_i = \lim_{s \to -\sigma_i} \frac{(s+\sigma_i)s^n}{\prod_{j=1}^{n} (s+\sigma_j)} \tag{6}$$

$\sigma_i < \sigma_j$ for $i < j$, and A is a positive real constant. The proof is given in Appendix.

Next, consider the other two forms shown in Figs. 2(c) and (d). The necessary and sufficient conditions for N_c and N_d can be best expressed on the admittance basis. These are:

For N_c:

$$y_{11} = y_{22} = A + \sum_{i=1}^{n} \frac{A|\rho_i|s}{(s+\sigma_i)} \tag{7}$$

$$-y_{12} = A + \sum_{i=1}^{n} (-1)^i \frac{A|\rho_i|s}{(s+\sigma_i)} \tag{8}$$

For N_d:

$$y_{11} = y_{22} = As + \sum_{i=1}^{n} \frac{A|\rho_i|s}{s + \sigma_i} \tag{9}$$

$$-y_{12} = As + \sum_{i=1}^{n} (-1)^{n-i+1} \frac{A|\rho_i|s}{s + \sigma_i} \tag{10}$$

where ρ_i in (7), (8) and ρ_i in (9), (10) are of the same form as (3) and (6), respectively.

The necessary and sufficient conditions for RL and LC cases can be obtained directly by RC-to-RL and RC-to-LC transformations.

III. NECESSARY AND SUFFICIENT CONDITIONS FOR THE REALIZABILITY OF NON-SERIES-PARALLEL BISYMMETRICAL TWO-PORTS

Consider the typical non-series-parallel bisymmetrical RC two-terminal-pair network (Fig. 3(a)). Bartlett's bisection theorem can be applied to this network [9]; its open-circuit-half and short-circuit-half are shown in Figs. 3(b) and (c), respectively. As shown in the last section, the necessary and sufficient condition for $z_{11(sh)}$, $z_{12(sh)}$ and $z_{22(sh)}{}^{*}$ to be realized as the symmetrical ladder network (Fig. 3(c)) is that they have the form of (1) and (2). From the Tee equivalent (Fig. 4) of the resultant ladder network of Fig. 3(c),

$$Z_{\frac{1}{2}\,\text{s.c.}} = z_{11} - z_{12} = z_{11(sh)} + z_{22(sh)} - 2z_{12(sh)} = \sum_{i=1,3,\dots} \frac{2A\,|\rho_i|}{s+\sigma_i} \tag{11}$$

$Z_{\frac{1}{2}\,\text{o.c.}}$ differs from $Z_{\frac{1}{2}\,\text{s.c.}}$ by a frequency scaling factor "a":

$$Z_{\frac{1}{2}\,\text{o.c.}} = z_{11} + z_{12} = \sum_{i=1,3,\dots} \frac{2A\,|\rho_i|}{as+\sigma_i} \tag{12}$$

Therefore, for a set of z_{ij} to be realizable as the non-series-parallel network in Fig. 3(a), z_{ij} must have the following forms

$$z_{11} = z_{22} = \sum_{i=1}^{n} \left[\frac{k_i}{as+\sigma_i} + \frac{k_i}{s+\sigma_i}\right] \tag{13}$$

$$z_{12} = \sum_{i=1}^{n} \left[\frac{k_i}{as+\sigma_i} - \frac{k_i}{s+\sigma_i}\right] \tag{14}$$

$$= \sum_{i=1}^{n} \left[\frac{(1-a)k_i s}{(as+\sigma_i)(s+\sigma_i)}\right] = \sum_{i=1}^{n} [f_i(s)] \tag{14'}$$

Theorem: The necessary and sufficient conditions for a set of RC z-parameters to be realized as the network of Fig. 3(a) is that z_{11}, z_{12}, and z_{22} satisfy (13) and (14).

The above analysis shows the necessity. The proof of sufficiency as well as the synthesis procedure will be given in the next section.

Notice that for n = 1, z_{12} only has a transmission zero at the origin. For n = 2, the transmission zeros may be either real or complex, but they all lie in the left half of the complex plane (LHP). Since the $f_i(s)$ are positive real functions and the sum of positive real function is a positive real function, for any n, z_{12} is a positive real function of s, or z_{12} is an RLC driving-point immittance function. Thus the transmission zeros of this class of functions all lie in LHP.

IV. A NON-SERIES-PARALLEL SYNTHESIS PROCEDURE (PROOF OF SUFFICIENCY)

The non-series-parallel synthesis procedure can be obtained as follows. Assume a given set of parameters z_{ij} which have the form (13) and (14). Then $Z_{\frac{1}{2}\text{ o.c.}}$ and $Z_{\frac{1}{2}\text{ s.c.}}$ are

$$Z_{\frac{1}{2}\text{ o.c.}} = \sum_{i=1}^{n} \frac{2k_i}{as + \sigma_i} \tag{15}$$

$$Z_{\frac{1}{2}\text{ s.c.}} = \sum_{i=1}^{n} \frac{2k_i}{s + \sigma_i} \tag{16}$$

Our previous discussion shows that if $Z_{\frac{1}{2}\text{ s.c.}}$ has the form in (16), $z_{ij(sh)}$ must be one of the following two forms:

Form I:

$$\begin{aligned} z_{11(sh)} = z_{22(sh)} = {} & \frac{A}{s} + \frac{A\,|\rho_1|}{s+\sigma_i} + \frac{A\,|\rho_{12}|}{s+\lambda_{12}} + \frac{A\,|\rho_2|}{s+\sigma_2} + \cdots + \frac{A\,|\rho_{n-1}|}{s+\sigma_{n-1}} \\ & + \frac{A\,|\rho_{n-1,n}|}{s+\lambda_{n-1,n}} + \frac{A\,|\rho_n|}{s+\sigma_n} \end{aligned} \tag{17}$$

$$\begin{aligned} z_{12(sh)} = {} & \frac{A}{s} - \frac{A\,|\rho_1|}{s+\sigma_1} + \frac{A\,|\rho_{12}|}{s+\lambda_{12}} - \frac{A\,|\rho_2|}{s+\sigma_2} + \cdots - \frac{A\,|\rho_{n-1}|}{s+\sigma_{n-1}} + \frac{A\,|\rho_{n-1,n}|}{s+\lambda_{n-1,n}} \\ & - \frac{A\,|\rho_n|}{s+\sigma_n} \end{aligned} \tag{18}$$

where, $\sigma_i < \lambda_{ij} < \sigma_j$,

$$\rho_i = \lim_{s\to -\sigma_i} \frac{(s+\sigma_i)\left[\prod_{p=1}^{n} \sigma_p\right]\cdot\left[\prod_{p=1}^{n-1} \lambda_{p,p+1}\right]}{s\cdot\left[\prod_{p=1}^{n} (s+\sigma_p)\right]\cdot\left[\prod_{p=1}^{n-1} (s+\lambda_{p,p+1})\right]} \tag{19}$$

$$\rho_{ij} = \lim_{s\to -\lambda_{i,j}} \frac{(s+\lambda_{i,j})\left[\prod_{p=1}^{n} \sigma_p\right]\cdot\left[\prod_{p=1}^{n-1} \lambda_{p,p+1}\right]}{s\cdot\left[\prod_{p=1}^{n} (s+\sigma_p)\right]\cdot\left[\prod_{p=1}^{n-1} (s+\lambda_{p,p+1})\right]} \tag{20}$$

Form II:

$$z_{11(sh)} = z_{22(sh)} = \frac{A}{s} + \frac{A|\rho_1|}{s+\sigma_i} + \frac{A|\rho_{12}|}{s+\lambda_{12}} + \frac{A|\rho_2|}{s+\sigma_2} + \cdots + \frac{A|\rho_{n-1}|}{s+\sigma_{n-1}}$$

$$+ \frac{A|\rho_{n-1,n}|}{s+\lambda_{n-1,n}} + \frac{A|\rho_n|}{s+\sigma_n} + \frac{A|\rho_{n,n+1}|}{s+\lambda_{n,n+1}} \tag{21}$$

$$z_{12(sh)} = \frac{A}{s} - \frac{A|\rho_1|}{s+\sigma_1} + \frac{A|\rho_{12}|}{s+\lambda_{12}} - \frac{A|\rho_2|}{s+\sigma_2} + \cdots - \frac{A|\rho_{n-1}|}{s+\sigma_{n-1}} + \frac{A|\rho_{n-1,n}|}{s+\lambda_{n-1,n}}$$

$$- \frac{A|\rho_n|}{s+\sigma_n} + \frac{A|\rho_{n,n+1}|}{s+\lambda_{n,n+1}} \tag{22}$$

where λ_{ij}, ρ_i, $\rho_{i,j}$ are defined as in (19), (20) except that ρ runs from 1 to n instead of from 1 to n - 1. Furthermore, the following relations must be held for all i, i = 1,...,n.

$$A|\rho_i| = k_i \tag{23}$$

For Form I, there are n linearly independent equations to be solved for n unknowns $(A, \lambda_{12}, \lambda_{23}, \ldots, \lambda_{n-1,n})$. For (17), (18) to be realizable as in Fig. 3(c), it is required that

(1) A and $\lambda_{j,j+1}$, j = 1,2,...,n - 1 are positive real

(2) $\sigma_i < \lambda_{i,i+1} < \sigma_{i+1}$ for all i.

Now one may ask the following questions: Is there such a solution which satisfies these two conditions? Do we need constraints on σ_i and k_i to obtain such a solution? One can verify that there always exists such a positive solution and there is no constraint on σ_i and k_i. Moreover, such a solution is unique for a given set of σ_i and k_i. When n is large, a computer is needed to solve $\lambda_{j,j+1}$, j = 1,2,...,n - 1.

For Form II, there are n + 1 unknowns $(A, \lambda_{12}, \lambda_{23}, \ldots, \lambda_{n,n+1})$ but only n equations. There is one degree of freedom. One can arbitrarily choose one of the unknowns so long as it does not violate the above two conditions. After doing so, we have n equations with n unknowns. The proof of the existence of a solution which satisfies the two conditions is similar to that for Form I.

In summary, for a given $Z_{\frac{1}{2}\,s.c.}$, as in (16), there always exists a solution of $z_{ij(sh)}$ of Form I (n < 1) which is unique and has infinitely many solutions of Form II. Therefore, the conditions on z_{11}, z_{12} and z_{22} in (13) and (14) are sufficient conditions for the realizability of the non-series-parallel network in Fig. 3(a). The synthesis procedure can now be summarized:

(1) Find $Z_{\frac{1}{2}\,s.c.}$ and $Z_{\frac{1}{2}\,o.c.}$ from the given z_{ij} which have the form of (13) and (14).

(2) Obtain $z_{ij(sh)}$ by solving A and λ_{ij} from (23).

(3) Realize $z_{ij(sh)}$ as a symmetrical ladder (Fig. 3(c)).

(4) Find the realization of $Z_{\frac{1}{2}\,s.c.}$ and $Z_{\frac{1}{2}\,o.c.}$, then the complete realization.

V. FURTHER NON-SERIES-PARALLEL REALIZATIONS

Next consider the network in Fig. 5(a). Observe that this network has electrical and structural bisymmetry. Bartlett's bisection theorem can be applied. The open-circuited half and short-circuited half are shown in Fig. 5(b) and (c), respectively. Notice that $Z_{\frac{1}{2}\,o.c.}$ is the same as z_{11} in Fig. 3(a), i.e.,

$$Z_{\frac{1}{2}\,o.c.} = \sum_{i=1}^{n} \left[\frac{k_i}{as + \sigma_i} + \frac{k_i}{s + \sigma_i} \right] \tag{24}$$

Moreover, it can be easily shown that $Z_{\frac{1}{2}\,s.c.}$ must be

$$Z_{\frac{1}{2}\,s.c.} = \sum_{i=1}^{n} \frac{2k_i}{(1+a)s + 2\sigma_i} \tag{25}$$

Therefore, the sufficient condition for a set of parameters z_{ij} to be realizable as the non-series-parallel network in Fig. 5(a) is that z_{ij} have the form

$$z_{11} = z_{22} = \sum_{i=1}^{n} \left[\frac{k_i}{as + \sigma_i} + \frac{k_i}{s + \sigma_i} \right] + \sum_{i=1}^{n} \frac{2k_i}{(1+a)s + 2\sigma_i} \tag{26}$$

$$z_{12} = \sum_{i=1}^{n} \left[\frac{k_i}{as + \sigma_i} + \frac{k_i}{s + \sigma_i} \right] - \sum_{i=1}^{n} \frac{2k_i}{(1+a)s + 2\sigma_i} \tag{27}$$

(26) and (27) are the necessary and sufficient conditions for a set of z-parameters to be realized as the network of Fig. 5(a).

One also can show that z_{12} of (27) is a positive real function for any n. The proof is similar to that of proving z_{12} of (14). Therefore, the transmission zeros of this class of functions all lie in LHP and may either be real or complex.

VI. EXAMPLE

The following simple example illustrates the synthesis procedure. Suppose that

$$z_{11} = z_{22} = \frac{1}{10s + 1} + \frac{1}{10s + 2} + \frac{1}{s + 1} + \frac{1}{s + 2} + \frac{2}{11s + 2} + \frac{2}{11s + 4} \tag{28}$$

$$z_{12} = \frac{1}{10s + 1} + \frac{1}{10s + 2} + \frac{1}{s + 1} + \frac{1}{s + 2} - \frac{2}{11s + 2} - \frac{2}{11s + 4} \tag{29}$$

It is seen that a = 10.
One can easily find that

$$Z_{\frac{1}{4}\,\text{o.c.}} = 2\left[\frac{2}{10s + 1} + \frac{2}{10s + 2}\right]$$

$$Z_{\frac{1}{4}\,\text{s.c.}} = 2\left[\frac{2}{s + 1} + \frac{2}{s + 2}\right]$$

where $k_1 = 2$, $k_2 = 2$.

(a) Form I

One finds that the two linearly independent equations are $A|\rho_i| = 2$ and $A|\rho_2| = 2$. Solving A and λ_{12} from the two equations, one obtains $\lambda = 5/3$ and $A = 2/5$. ρ_1, ρ_2 and ρ_{12} are found to be*

$$\rho_1 = -5\,, \quad \rho_{12} = 9 \quad \text{and} \quad \rho_2 = -5\,.$$

The open-circuited half and the short-circuited half of this symmetrical ladder denoted by $[Z_{\frac{1}{2}\,\text{o.c.}}]_L$ and $[Z_{\frac{1}{2}\,\text{o.c.}}]_L$, respectively, are

$$[Z_{\frac{1}{2}\,\text{o.c.}}]_L = z_{11(sh)} + z_{12(sh)} = \frac{8s + \frac{4}{3}}{s^2 + \frac{5}{3}s}$$

$$[Z_{\frac{1}{2}\,\text{s.c.}}]_L = z_{11(sh)} - z_{12(sh)} = \frac{8s + 12}{s^2 + 3s + 2}$$

The realizations are shown in Fig. 6(a) and (b). The symmetrical ladder which realizes $z_{ij(sh)}$ is given in Fig. 6(c).

(b) Form II

One has two equations: $A|\rho_1| = 2$ and $A|\rho_2| = 2$, but there are three unknowns A, λ_{12}, and λ_{23}. Thus one needs to set one of them equal to a constant, say $\lambda_{23} = 3$. Then $\lambda_{12} = 3/2$ and $A = 2/9$. So, ρ_1, ρ_2, ρ_{12} and ρ_{23} are are $\rho_1 = -9$, $\rho_{12} = 16$, $\rho_2 = -9$, and $\rho_{23} = 1$. Thus

$$[Z_{\frac{1}{2}\,\text{o.c.}}]_L = z_{11(sh)} + z_{12(sh)} = \frac{8s^2 + 24s + 2}{s^3 + \frac{9}{2}s^2 + \frac{9}{2}s}$$

---

* It can be shown that $\Sigma\rho_i + \Sigma\rho_{ij} + 1 = 0$. This is a useful relation for checking purposes.

$$[Z_{\frac{1}{2}\text{ s.c.}}]_L = z_{11(sh)} - z_{12(sh)} = \frac{8s + 12}{s^2 + 3s + 2}$$

The realization of $[Z_{\frac{1}{2}\text{ o.c.}}]_L$ is shown in Fig. 7(a) and the realization of $[Z_{\frac{1}{2}\text{ s.c.}}]_L$ is the same as the realization of $[Z_{\frac{1}{2}\text{ o.c.}}]_L$ in Form I. The symmetrical ladder is given in Fig. 7(b).

(4) Find the complete realization. See Fig. 8.

VII. CONCLUSION

A non-series-parallel synthesis method has been given to realize a class of bisymmetrical RC (or RL or LC) two-terminal-pair network functions whose transmission zeros all lie in the LHP of the complex plane including the imaginary axis. By employing the same general approach which has just been described, and slightly varying the details, a number of related synthesis procedures can be developed:

(1) for the network which results when the resistors and capacitors (of Fig. 3(a)) are interchanged;

(2) for the network shown in Fig. 9 on an admittance basis;

(3) for the network which results when the resistors and capacitors of Fig. 9 are interchanged.

Additionally, another class of bisymmetrical networks exists to which the method given in this paper can be applied. A block diagram of a potentially bisymmetrical network is given in Fig. 10. The non-series-parallel synthesis procedure is readily generalized to the potentially bisymmetrical networks of the four classes of networks, just described.

REFERENCES

1. Kim, W. H., "A New Method of Driving Point Function Synthesis," Ph.D. Thesis, 1956, University of Illinois.

2. Van Valkenburg, M. E., "A Special Case of a Bridge Equivalent of Brune Networks," Proc. IRE, Vol. 44, November 1956, p. 1621.

3. Lee, S. C. and Frisch, I. T., "A Class of RLC Networks with Fewer Non-reactive Elements than the Brune Realization, IEEE Trans. on Circuit Theory, Vol. CT-11, No. 3, September, 1964, pp. 418-421.

4. Lee, S. C., "Network Synthesis to Minimize Multiparameter Sensitivity," Ph.D. Thesis, University of Illinois, August, 1965.

5. Cauer, W., Synthesis of Linear Communication Networks, McGraw-Hill Book Co., 1958, Chapter 5.

6. Van Valkenburg, M. E., Introduction to Modern Network Synthesis, John Wiley & Sons, Inc., 1962.

7. Fialkow, A. D. and Gerst, I., "The Transfer Function of General Two-Terminal-Pair RC Networks," Quart. Appl. Math., 10, 113-127 (1952).

8. Ozaki, H., "Synthesis of RC 3-Terminal Networks Without Ideal Transformers," Osaka Univ. Tech. Rept. 3, (60) pp. 55-77, 1953.

9. Friedland, B., Wing, O. and Ash, R., Principles of Linear Networks, McGraw-Hill Book Company, Inc., 1961, pp. 256-259.

APPENDIX

Proof of the Necessary and Sufficient Conditions for a Set of RC z_{11}, z_{12}, and z_{22} Parameters to be Realized as an Electrically and Structurally Symmetrical Canonical Ladder Network as Shown in Figs. 2(a) and (b)

Since the proofs of (a) and (b) are similar, there is no need to give both. The proof of (a) is given as follows:

Proof of Necessity: From N_a, one finds that z_{11}, z_{22} have the form

$$z_{11} = z_{22} = \frac{A}{s} + \sum_{i=1}^{n} \alpha_i \frac{1}{s + \sigma_i} \tag{A-1}$$

where A is a positive real constant and z_{12} may be expressed by

$$z_{12} = \frac{A}{s} + \sum_{i=1}^{n} (-1)^i A \,|\, \rho_i \,|\, \frac{1}{s + \sigma_i} \tag{A-2}$$

where ρ_i is defined as in (3).

Since N_a is a canonical realization, z_{ij} must be compact at all poles. This leads to $\alpha_i = A \,|\, \rho_i \,|$ for all i.

Proof of Sufficiency: Assume that the given z_{ij} satisfying the condition stated in (a) can be synthesized by N_a which is an electrically and structurally symmetrical network. Then Bartlett's bisection theorem can be applied. It is found that

$$Z_{\frac{1}{2}\,\text{o.c.}} = z_{11} + z_{12} = \frac{2A}{s} + \sum_{i=2,4,\ldots} \frac{2A \,|\, \rho_i \,|}{s + \sigma_i} \tag{A-3}$$

$$Z_{\frac{1}{2}\,\text{s.c.}} = z_{11} - z_{12} = \sum_{i=1,3,\ldots} \frac{2A \,|\, \rho_i \,|}{s + \sigma_i} \tag{A-4}$$

One finds that $Z_{\frac{1}{2}\,\text{o.c.}}$ and $Z_{\frac{1}{2}\,\text{s.c.}}$ are always consistent with Fig. 1(a). This means that our hypothesis is true. In fact, such a symmetrical canonical ladder network N_a can always be obtained. This completes the proof.

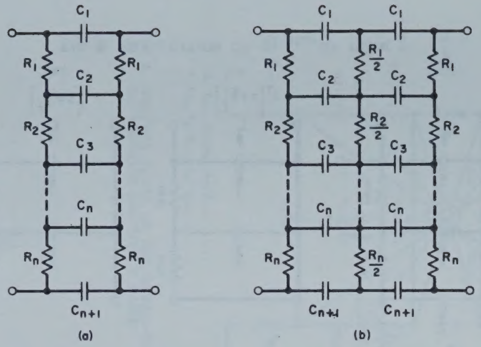
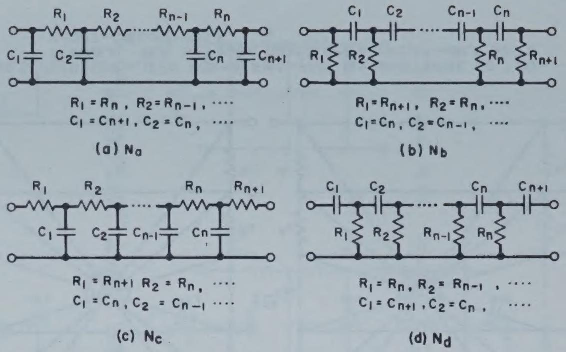


FIG. 1 NON-SERIES-PARALLEL REALIZATION OF SYMMETRICAL RC TWO-PORTS



$R_1 = R_n$, $R_2 = R_{n-1}$, ...
$C_1 = C_{n+1}$, $C_2 = C_n$, ...

(a) N_a

$R_1 = R_{n+1}$, $R_2 = R_n$, ...
$C_1 = C_n$, $C_2 = C_{n-1}$, ...

(b) N_b

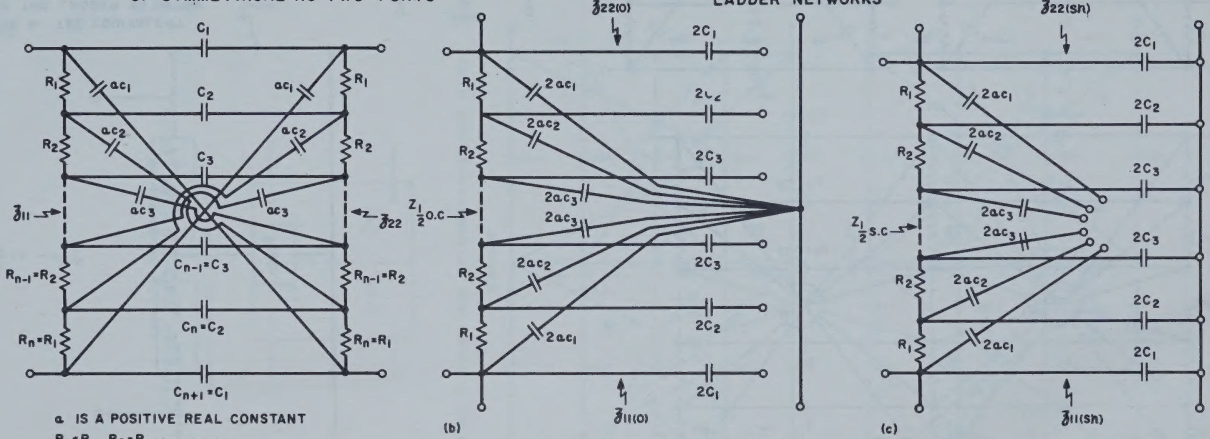
$R_1 = R_{n+1}$ $R_2 = R_n$, ...
$C_1 = C_n$, $C_2 = C_{n-1}$ ....

(c) N_c

$R_1 = R_n$, $R_2 = R_{n-1}$, ...
$C_1 = C_{n+1}$, $C_2 = C_n$, ...

(d) N_d

FIG. 2 ELECTRICALLY AND STRUCTURALLY SYMMETRICAL RC LADDER NETWORKS



a IS A POSITIVE REAL CONSTANT

$R_1 = R_n$, $R_2 = R_{n-1}$, ----
$C_1 = C_{n+1}$, $C_2 = C_n$, ----

(a)

FIG. 3 (a) NON-SERIES-PARALLEL BISYMMETRICAL RC TWO-TERMINAL-PAIR NETWORK (b) ITS OPEN-CIRCUIT-HALF AND (c) ITS SHORT-CIRCUIT-HALF

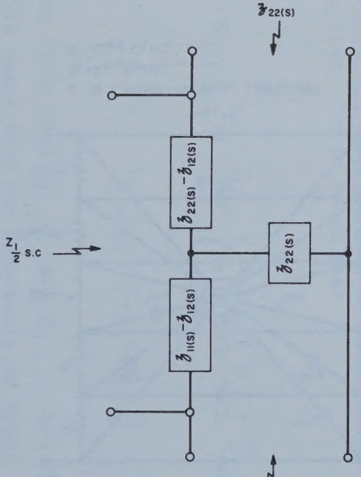

$z_{22(s)}$
$Z_{\frac{1}{2}}$ s.c
$z_{22(s)} - z_{12(s)}$
$z_{22(s)}$
$z_{11(s)} - z_{12(s)}$
$z_{11(sh)}$

FIG. 4 TEE EQUIVALENT OF THE LADDER NETWORK OF FIG. 3(c)

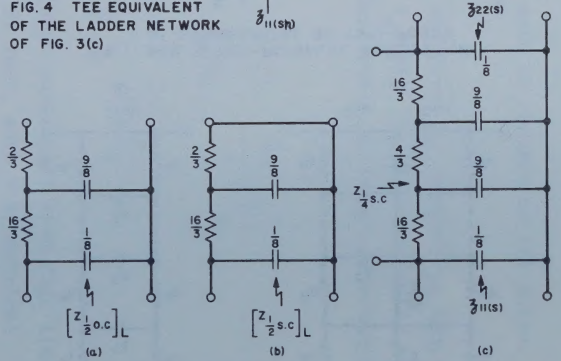

$\frac{2}{3}$
$\frac{16}{3}$
$\frac{9}{8}$
$\frac{1}{8}$
$\left[Z_{\frac{1}{2}\text{o.c}}\right]_L$
(a)
$\frac{2}{3}$
$\frac{16}{3}$
$\frac{9}{8}$
$\frac{1}{8}$
$\left[Z_{\frac{1}{2}\text{s.c}}\right]_L$
(b)
$z_{22(s)}$
$\frac{16}{3}$
$\frac{4}{3}$
$\frac{16}{3}$
$\frac{1}{8}$
$\frac{9}{8}$
$\frac{9}{8}$
$\frac{1}{8}$
$Z_{\frac{1}{4}}$ s.c
$z_{11(s)}$
(c)

FIG. 6 REALIZATION OF $z_{ij(sh)}$ OF FORM I

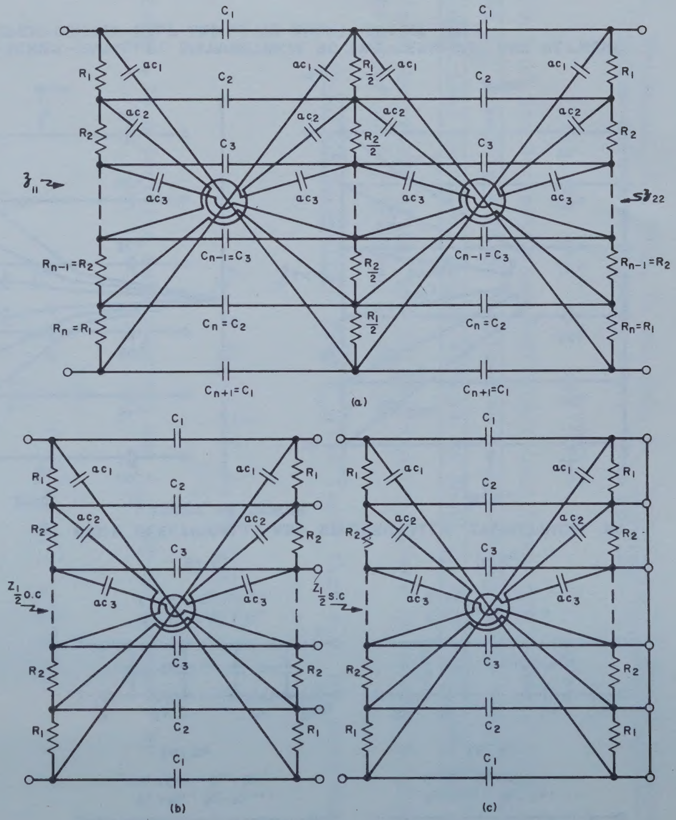

C_1
C_2
C_3
$C_{n-1}=C_3$
$C_n=C_2$
$C_{n+1}=C_1$
R_1
R_2
$R_{n-1}=R_2$
$R_n=R_1$
$\frac{R_1}{2}$
$\frac{R_2}{2}$
ac_1
ac_2
ac_3
z_{11}
z_{22}
(a)
$Z_{\frac{1}{2}\text{o.c}}$
(b)
$Z_{\frac{1}{2}\text{s.c}}$
(c)

FIG. 5 (a) GENERALIZED NON-SERIES-PAIR BISYMMETRICAL RC TWO-TERMINAL-PAIR NETWORK (b) ITS OPEN-CIRCUIT-HALF (c) ITS SHORT-CIRCUIT-HALF.

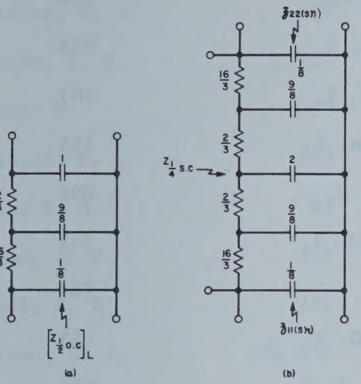
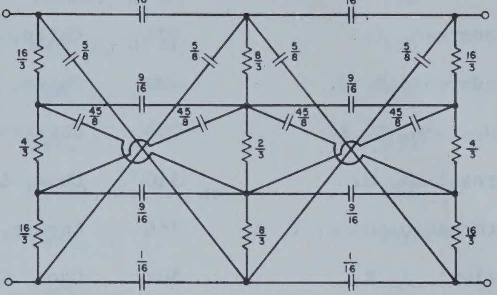
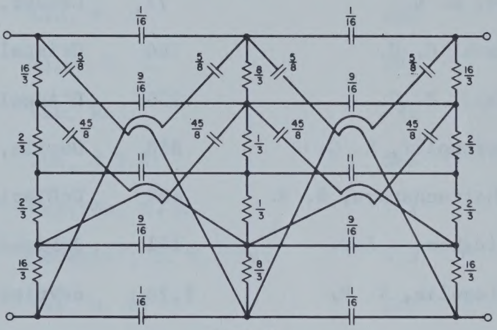

2/3
16/3
1
9/8
1/8
$\left[Z_{\frac{1}{2}\,o.c}\right]_L$
(a)
$\mathfrak{z}_{22}(sh)$
1/8
16/3
9/8
$Z_{\frac{1}{4}\,s.c}$
2/3
2
2/3
9/8
16/3
1/8
$\mathfrak{z}_{11}(sh)$
(b)

FIG. 7 REALIZATION OF $\mathfrak{z}_{11}(sh)$ OF FORM II


1/16
1/16
16/3
5/8
5/8
8/3
5/8
5/8
16/3
9/16
9/16
45/8
45/8
45/8
45/8
4/3
2/3
4/3
9/16
9/16
16/3
8/3
16/3
1/16
1/16

(a) FORM I


1/16
1/16
16/3
5/8
5/8
8/3
5/8
5/8
16/3
9/16
9/16
45/8
45/8
1/3
45/8
45/8
2/3
11
11
2/3
1/3
2/3
2/3
9/16
9/16
16/3
8/3
16/3
1/16
1/16

(b) FORM II

FIG. 8 NON-SERIES-PARALLEL REALIZATION OF (28) (29)

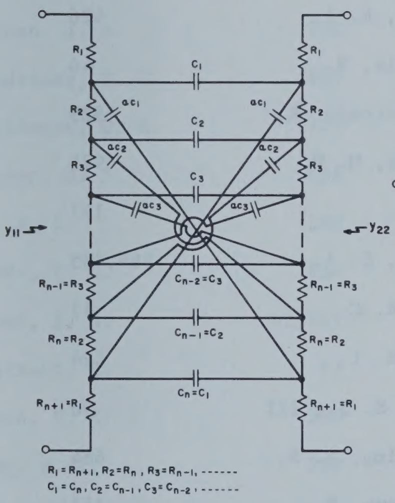
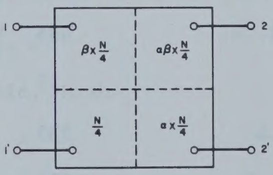

R₁
C₁
R₁
R₂
αc₁
C₂
αc₁
R₂
αc₂
αc₂
R₃
C₃
R₃
αc₃
αc₃
y₁₁
y₂₂
$R_{n-1} = R_3$
$C_{n-2} = C_3$
$R_{n-1} = R_3$
$R_n = R_2$
$C_{n-1} = C_2$
$R_n = R_2$
$C_n = C_1$
$R_{n+1} = R_1$
$R_{n+1} = R_1$
$R_1 = R_{n+1}$, $R_2 = R_n$, $R_3 = R_{n-1}$, ------
$C_1 = C_n$, $C_2 = C_{n-1}$, $C_3 = C_{n-2}$, ------

FIG. 9 ANOTHER CLASS OF NON-SERIES-PARALLEL NETWORKS (ON ADMITTANCE BASIS)


1
2
$\beta \times \frac{N}{4}$
$\alpha\beta \times \frac{N}{4}$
$\frac{N}{4}$
$\alpha \times \frac{N}{4}$
1'
2'

α, β ARE MAGNITUDE SCALING FACTORS

FIG. 10 A BLOCK DIAGRAM OF POTENTIALLY BISYMMETRICAL NETWORK

INDEX OF AUTHORS

Adams, J. V. 708

Andersen, J. 934

Anderson, B. D. 400

Anderson, D. R. 708

Arozullah, M. 310

Athanassopoulos, I. A. 766

Atluri, S. R. 508

Au, S. K. 71

Bach, G. G. 186

Ball, E. C. 250

Bertnolli, E. C. 868

Bhattacharyya, B. B. 488

Bingham, J.A.C. 141

Bingulac, S. P. 8,28

Biswas, R. N. 282

Blank, G. L. 869

Brockett, R. W. 685

Burzio, A. 17

Calvin, E. L. 335

Carlson, A. J. 334

Carpenter, R. M. 454

Chan, S. P. 227

Chen, C. T. 785

Chen, W. H. 933

Chen, W. K. 207

Chen, Y. K. 868

Chien, R. T. 359

Chien, Y. T. 107

Christian, E. 155

Chua, L. O. 698

Chyung, D. H. 587

Cook, G. 567

Cooper, G. R. 53

Cristal, E. G. 858

D'Angelo, H. 489

Davila, E. A. 426

DeClaris, N. 46

deFigueiredo, R. 925

deVries, M. M. 925

Deo, N. 197

Desoer, C. A. 669,785

Dorf, R. C. 17

Duty, R. L. 684

Dwyer, S. J., III 96

Eggerling, J. B. 684

Eisenmann, E. 155

Etzweiler, G. A. 548

Fountain, G. H. 914

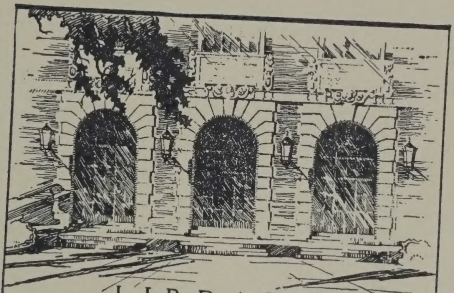
Franks, L. E. 649

Fu, K. S. 107,794

Funk, J. E. 567
Glinski, G. S. 310
Gonen, B. 165
Gonzales, R. L. 527
Gopalan, R. S. 731
Gruber, M. 827
Gupta, S. C. 849
Haase, K. H. 83
Haddad, A. H. 322
Hakimi, S. L. 197
Halijak, C. A. 868,914
Hammond, W. M. 53
Happ, W. W. 454,849
Heinen, J. A. 751
Hendricks, T. C. 489
Heylinger, G. E. 175
Hinden, H. J. 486
Ho, B. L. 388
Hooks, C. G. 684
Hoyer, E. A. 82,249
Ishizaki, Y. 61
Jones, H. E. 259
Jong, M. T. 237
Kahne, S. J. 499
Kalman, R. E. 388
Kleckner, K. R. 659
Kokotović, P. 377
Kriger, L. V. 474
Kwan, R. K. 186
Lago, G. V. 82
Larson, R. E. 600
Lasdon, L. S. 753
Layton, D. M. 639
Lee, H. B. 934
Lee, S. C. 269,955
Leeds, J. V., Jr. 466
Leon, B. J. 630
Levy, L. J. 869
Lin, P. M. 630
Liou, M. L. 291
Liu, R. 71
Luh, J.Y.S. 549
McBride, L. M. 741
McDaniel, L. W., Jr. 913
Mai, H. N. 227
Marshall, T. G., Jr. 935
Meadows, H. E. 649,776
Medanić, J. 377
Mesarovic, M. D. 1
Michel, A. N. 752
Milic, M. M. 218
Mitter, S. K. 753
Moe, M. L. 489
Moore, J. 887
Muller, D. E. 368
Mullick, S. K. 674

Murata, T. 244
Myers, B. R. 408,416
Narendra, K. S. 814
Navot, I. 165
Neuman, C. P. 814
Newcomb, R. W. 669
O'Donnell, J. J. 577
O'Neill, J. F. 72
Olivares, J. E., Jr. 944
Orchard, H. J. 465
Parker, S. R. 538
Parten, M. 487
Pellegrini, C. A. 877
Perdikaris, G. 249
Peri, Z. 165
Porter, W. A. 517
Preparata, F. P. 359
Pullen, K. A., Jr. 726
Puri, N. N. 903
Ramapriyan, H. K. 915
Reddy, D. C. 283
Resh, J. A. 621
Rogers, J. W. 70
Rutman, R. 377
Sain, M. K. 593
Schoeffler, J. D. 116,766
Schoenberger, M. 557
Schultz, D. G. 916
Schwarzlander, H. 805
Seacat, R. H. 487
Shafran, J. S. 549
Shapiro, S. D. 352
Shenoi, B. A. 259,731
Shuraym, G. 485
Siljak, D. 887
Silverman, L. M. 649,776
Simes, J. G. 548
Sinha, N. K. 508
Souza, C. 611
Srinath, M. D. 915
Steele, S. A. 876
Steiglitz, K. 741
Stern, T. E. 819
Swain, P. H. 794
Swamy, M.N.S. 488
Takahara, Y. 1
Tang, D. T. 427
Tapia, M. A. 416
Taub, J. J. 486
Temes, G. C. 465
Thaler, G. J. 250,877
Thathachar, M.A.L. 915
Thomas, J. B. 322
Trick, T. N. 718
Tsuchiya, T. 61
Van Landingham, H. 46
Van Mellaert, L. J. 300
Waag, R. C. 659

Wade, W. D.	334
Wagner, T. J.	90
Waltz, F. M.	36
Waren, A. D.	116,753,766
Wasta, R.	876
Weinberg, L.	433,444
Weygandt, C. N.	903
Whitney, A. W.	96
Willems, J. C.	836
Willems, J. L.	827
Wong, K. K.	669
Wozny, M. J.	916
Wu, S. H.	751,752
Yau, S. S.	343
Yockey, H. M.	877
Yoshida, N.	61
Zobrist, G. W.	237,249



LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
4 1966
3 0112 006279613